Empirical models of tree top breakage and scattered wind-throw in a *Pinus radiata* and *Pseudotsuga menziesii* plantation forest in New Zealand

A thesis
submitted in fulfilment
of the requirements for the Degree
of
Master of Forestry Science
at
the University of Canterbury
by
Yannina Whiteley

January 2020

Acknowledgements

I thank my academic supervisors, Dr. Justin Morgenroth and Dr. Varvara Vetrova, for all their assistance. Many thanks also to Port Blakely Ltd, who provided the data for this research, and to Mr. Aaron Gunn of Port Blakely, who patiently answered many questions, especially during the project formulation stages. I owe much to Dr. Serajis Salekin, who gave generously of his time to review each and every chapter, asking nothing but in return but appreciation. Last, but certainly not least, I acknowledge my husband Simon, who made the time available in our lives for me to pursue this goal.

Table of Contents

List of tables

List of figures

Abstract

In this thesis, attritional damage to trees attributed to wind and snow was studied in *Pinus radiata* (radiata pine) and *Pseudotsuga menziesii* (Douglas-fir) at Geraldine Forest, a 5,500 hectare forest in the South Island of New Zealand, based on tree damage descriptions from forest sample plot data. This damage largely comprises the breakage of live stems, and also a smaller number of windthrown trees. The mean levels of damage were compared by species, and it was established that the damage levels are significantly different, with Douglas-fir exhibiting higher broken heights, lower proportion of trees damaged, and lower proportions of trees alive than radiata pine.

With these results established, damage was modelled for each species separately, using both mixed-effects linear (or generalised linear) regression and random forests to create empirical models. The three response variables were the mean height of broken trees per plot (*P_tree_ht_mean_BRKN*); the proportion of trees with damaged tops per plot (*Tops_prpn_DAM*); and the proportion of live trees per plot (*Prpn_LIVE*). None of the models created were suitable for re-use with new data, due to bias in the model results and the reliance of the models on mixed-effects.

Three models had sufficient explanatory power to demonstrate that some particular tree and topographic variables correlate with damage levels. These models, all at the plot level, were 1) radiata pine *P_tree_ht_mean_BRKN* by linear regression with mixed-effects, 2) radiata pine *P_tree_ht_mean_BRKN* by random forest, and 3) radiata pine *Tops_prpn_DAM* by logistic regression with mixed-effects, using only data where the top of every tree had been assessed.

For the model of radiata pine *P_tree_ht_mean_BRKN* by linear regression with mixed-effects, the age of the trees, the proportion of trees pruned, and the aspect correlated with the height at which trees break, with the stand identity as a mixed effect. For the model of radiata *P_tree_ht_mean_BRKN* by random forests, the year of establishment, the per-hectare equivalent basal area, the age of the trees, the height of the unbroken trees, the pruned proportion, the mean pruned height, and a group of weather variables correlated with the height at which trees break. For the model of radiata pine *Tops_prpn_DAM* by generalised linear model with mixed-effects, the per-hectare equivalent stocking, the mean diameter of unbroken trees, and the proportion of live trees correlated with the proportion of trees damaged, with the stand identity and the plot number as mixed effects.

From these results, and by comparison with previous research into empirical models of damage to trees by wind and snow, some management recommendations have been made to reduce future damage by wind and snow at Geraldine Forest. The first is that if low levels of damage are highly desired, then Douglas-fir is the better species to plant. To reduce the levels of damage in radiata pine, any or all of the following measures apply. The first is to avoid growing radiata pine on slopes with north-east and or south-east aspects, and/or in areas of low topographic shelter, both of which positively correlate with higher proportion damaged. The second is to choose a low stocking for radiata pine, as high stocking is correlated with higher proportion damaged, but without implementing very heavy or very late thinning. The third recommendation is short rotations for radiata pine, as the age of trees is a strong predictor of damage levels. The fourth recommendation is to plant radiata pine at low elevations; height growth is faster at lower elevations and so trees will attain a desirable size in a shorter rotation; also, taller trees have higher broken heights, leaving a longer salvageable portion of stem below any breaks. The fifth recommendation (which runs somewhat counter to the third) is to prune the radiata pine crop, because pruned radiata pine breaks at higher heights, again leaving more salvageable stem.

1 Introduction

1.1 Origins of the research topic

Geraldine Forest, in the Timaru District of Canterbury, New Zealand, is a timber-producing plantation forest comprising mostly two evergreen conifers: *Pinus radiata (D. Don)*, commonly called radiata pine, and *Pseudotsuga menziesii ((Mirb.) Franco)*, commonly called Douglas-fir. Geraldine Forest has a high breakage rate for standing radiata pine. Windthrown trees also occur, often intermingled with standing broken trees. Collectively referred to in this research as tree damage, standing tree breakage and windthrown trees at Geraldine Forest cause substantial and spatially variable volume losses and downgrade of stems, to the point of adversely affecting forest management and forest profitability.

Port Blakely's New Zealand Forestry division owns Geraldine Forest. Port Blakely staff suspect that standing tree breakage and scattered windthrow occurs mostly during winter storms that deliver either wind or both wind and snow. Wind plus snow appears to have the largest effects on stands of age 3 to 10 years, whereas wind-alone event affect older-aged stands. Staff further suspect that variability in standing tree breakage relates to some or all of tree species, site biomass, elevation, tree age, storm snowfall amount, storm wind direction and strength, and topographic factors, particularly lee slopes with regard to the dominant wind direction of storms.

Port Blakely would like a better understanding of standing tree breakage and windthrow at Geraldine Forest, for estimation of probable losses within stands, and also to devise risk management strategies.

1.2 The importance of forestry in New Zealand

New Zealand has a substantial proportion of its 26.8 million hectares in forest. Forests in New Zealand fall primarily into two groups. The first is indigenous forests and related vegetation types, totalling approximately 8.7 million hectares, which are largely in reserves and are not subject to management for timber production (Landcare Research New Zealand Ltd, 2015). The second is plantation forests, which are highly managed for timber production, which total approximately 1.72 million hectares and include 1.55 million hectares of radiata pine, with Douglas-fir as the next most common species (Ministry for Primary Industries, 2018). Geraldine Forest is part of this second group. Plantation forestry is a substantial contributor to New Zealand's economy, contributing $1,389,000 per year, or 0.6% of gross domestic product (Nixon, Gamperle, Pambudi, & Clough, 2017).

1.3 Aspects of this research

1.3.1 Research goals

The goals for this research are:

- to show that quality scientific research can be undertaken from publically-available data in conjunction with data that New Zealand forest managers often collect in the pursuit of non-research goals, such as yield estimation.
- to discern whether there are differences in damage from wind and snow for radiata pine and Douglas-fir at Geraldine Forest, where they are the two main species planted.

- to identify and understand factors that correlate with wind and snow damage to trees at Geraldine Forest, especially damage-worsening factors that may be mitigated against by forest management practices.
- to create models to explain that damage, preferably models which can also be used to make sound predictions from new data.

1.3.2 Research questions

These research goals gave rise to a set of research questions:

1. Do rates of tree damage differ significantly for radiata pine and Douglas-fir at Geraldine Forest?
2. How well can tree damage in radiata pine and Douglas-fir at Geraldine Forest be modelled?
3. Which modelling approach, regression or random forests, creates the most explanatory and least biased models?
4. Which tree and stand characteristics and which topographic conditions significantly affect tree damage at Geraldine Forest?
5. Can the models developed be used to predict damage from new data?
6. Do the research findings suggest forest management practices that may reduce tree damage in radiata pine at Geraldine Forest?

These research questions will be answered by:

- Statistical analysis to determine differences in tree damage between radiata pine and Douglas-fir.
- Identification of a set of explanatory variables with potential for use in modelling standing tree breakage and associated windthrow.
- For each species, development of models that detail the height at which trees break in sampling plots.
- For each species, development of models that detail the proportion of trees damaged in sampling plots.
- For each species, development of a model of the proportion of live trees in sampling plots.
- Validation of all models created.
- Using model outcomes and the findings from previous literature to make suggestions for forest management practices that may reduce the risk of damage.

1.3.3 Research benefits

Measured timber yields for Geraldine Forest from harvesting, at around age 28, average 15% lower than the figures indicated by yield models initiated at age 18 for the same trees. This compares with only an 8% drop in two similar forests owned by Port Blakely, which indicates that the volume drop at Geraldine Forest is due to tree damage over that period, additional to any yield loss from suboptimal harvesting practices. The 2016 LiDAR survey of this forest showed 40% of *Pinus radiata* trees 14 years and older had broken tops[1]. Therefore, this research will benefit the forest owner by providing a better understanding of standing tree breakage and scattered windthrow at Geraldine Forest, which may

[1] A. Gunn, Port Blakely, pers. comm.

allow for site-tailored future management to reduce risk. More generally, this research is intended to demonstrate a technique for modelling and assessing the risk of dispersed wind or wind plus snow damage in any plantation where a similar information base exists.

1.3.4 Thesis outline

This chapter, Chapter One, introduces the research topic and the relevant literature. Chapter Two describes the methodology used to gather and process forest data, to calculate topographic and weather data, and to create models. Results are presented in Chapter Three. Chapter Four summarises and discusses the major results. Chapter Five concludes the thesis by considering the implications of the results for the management of radiata pine and Douglas-fir at Geraldine Forest. Chapter Six contains appendices that support the material presented in the first five chapters.

1.4 Literature review

1.4.1 A brief background to wind and snow damage

New Zealand plantation forests have a significant history of wind damage. An overview of wind damage by Somerville (1995) compiled records of damage back to the 1940s; at least 50,000 ha had been catastrophically damaged to the point where it could no longer be considered manageable forest, with an additional unknown amount of attritional damage to remaining forest. Park (2009) updated the damage estimate to 60,000 ha. Moore, Manley, Park, and Scarrott (2013), used wind damage records as a basis for the calculations of past losses and possible future losses per unit time, at the forest level and at the wood supply region level. Wind damage to New Zealand has been the subject of workshops and articles drawing conclusions and management recommendations from observations of wind events, sometimes specific wind events; for example, see Somerville, Wakelin, and Whitehouse (1989): *Workshop on Wind Damage in New Zealand exotic forests*. The only published study discovered that involves wind and snow damage at Geraldine Forest in particular is that of Ledgard (1982), who discusses some wind damage history and possible relationships to topography in their context as factors that may influence the outcomes of silvicultural regime analysis.

Many authors have drawn a distinction between wind damage affecting large areas and windthrow affecting small areas or single trees. The split is variously termed catastrophic versus attritional (Somerville, 1995), catastrophic versus chronic (Everham & Nicholas, 1996), coarse-scale versus fine-scale (Rebertus, Kitzberger, Veblen, & Roovers, 1997), catastrophic versus background (Rifai et al., 2016), with accompanying concept that the return interval of the catastrophic events is much longer than the return interval of the attrition-causing events. This tendency to make a distinction was criticised by Mitchell (2013), who felt that windthrow was too often considered an exception rather than a recurrent natural disturbance. Nevertheless, this study imposes the distinction of scattered versus widespread damage, and deals with only the former.

1.4.2 Types of wind and snow damage studies

Schindler, Bauhus, and Mayer (2012), in their review of wind effect on trees, usefully group modelling techniques into investigations of the causes of damage, which may be statistical, semi-mechanistic or mechanistic; and empirical statistical models of the probability of damage, which provide only general insights into causes. There has been research into development of mechanistic models of windthrow risk in New Zealand (Moore and Somerville (1998); Moore and Quine (2000)), and the use of such models (Moore and Gardiner (2001)), but little New Zealand work regarding empirical statistical models of wind or wind plus snow damage was located during this review. The most similar study located in the New Zealand literature is by Wrathall (1989) who studied the catastrophic effects of an ex-cyclonic windstorm (Cyclone Bola) on 17 – 23 year old radiata pine trees at Waitahanui Forest in the Central North Island. The international literature is more extensive and provides valuable insights from other areas with single-species conifer stands, particularly in Europe.

On a world-wide basis, as noted by Albrecht, Hanewinkel, Bauhus, and Kohnle (2012), most studies in the field of post-disturbance forest analysis take an empirical statistical approach based on observational data, as does this research. Many of the international individual studies discussed here follow single storm events considered severe or catastrophic; although this research involves scattered damage, the studies of catastrophic damage remain informative. Taking these points into consideration, along with the climate of and species planted at Geraldine Forest, a reasonably narrow view has been taken of relevant literature, limited to examining the effects on trees of wind, snow, and ice, with an emphasis on non-tropical coniferous forests, and an emphasis on empirical regression techniques.

1.4.3 Empirical studies of wind and snow damage

Empirical studies of wind and snow damage may be placed into three basic groups, and summaries of studies are presented in that fashion in Table 1-1 to Table 1-3, below. The first is analyses that use statistical modelling without an explicit spatial component. The second is analyses that use statistical models with topographic (i.e. spatial) variables amongst the explanatory variables, but no make explicit allowance for spatial relationships in the model equations; some of these studies test for spatial autocorrelation in their model residuals and some do not. The third group incorporates spatial relationships into the model equations (for example, by spatial regression), and may also include topographic variables among the explanatory variables. This research belongs to the second group, and includes tests for autocorrelation.

As well as quickly summarising the location, forest type, and statistical coverage of past empirical studies of wind and snow damage, another use of Table 1-1 to Table 1-3 is to briefly present important predictors of tree damage. These predictors, further discussed and elaborated on in section 1.4.4: *Explanatory variables in empirical wind and snow damage studies*, were influential in the development of the potential explantory variable set for this study. In this table, dec. = decreasing, inc. = increasing, sph = stems/ha, b.a. = basal area, dbh = diameter at breast height, *DA* = deciduous angiosperms, and modelling is by field survey, unless otherwise specified.

Table 1-1: empirical studies of wind and snow damage in the literature: modelling without explicit spatial component

study	damage by	modelling	species or forest type	location	important predictors of inc. damage
Wrathall (1989)	wind	chi-square tests for numbers of tree by different damage types, linear regression of logit-transformed probabilities of different damage types	*Pinus radiata*	Central North Island, New Zealand (the only NZ study in this table)	windthrown yes/no: high dbh; high taper; inc. years since thin broken yes/no: inc. age; inc. years since thin height of break & break as % tree length: inc. height, inc. dbh^2, inc. crown size
Munishi and Chamshama (1994)	wind	linear regression of % per transect breaking, bending or uprooting	*Pinus patula*	Southern Highlands, Tanzania	high height/dbh ratio; high sph
Fridman, Valinger, and Sveriges (1998)	wind	logistic regression of per-plot presence/ absence damage	*Pinus sylvestris*	Västerbotten, Sweden	high mean height; high volume index
Veblen, Kulakowski, Eisenhart, and Baker (2001)	wind	chi-squared tests & analysis of variance of patches of windthrow defined on aerial photos & classified into severity groups	*Populus tremuloides*; *Pinus contorta*; *Abies lasiocarpa*; *Picea engelmanii*	Colorado, U.S.A	species; high tree height; tree live/dead
Dobbertin (2002)	wind	cross-validated classification trees of damaged/not damaged status of post-storm inventory plots	many species	all of Switzerland	high stand height; older stands; high % of conifers; high soil water-logging; inc. soil depth
Scott and Mitchell (2005)	wind	logistic regression of damaged/not damaged status of trees in post-storm sample plots	*Tsuga heterophylla*; *Abies amabilis*	British Columbia, Canada	high height/dbh ratio; high crown density; dec. crown length; dec. tree retention; inc. wind run
Aubrey, Coleman, and Coyle (2007)	ice storm	analysis of variance of different proportions of damage between plots in an irrigation & fertilisation trial	*Pinus taeda*	South Carolina, U.S.A	high tree height; large dbh; high taper; high leaf, branch & crown biomass
Valinger and Fridman (2011)	wind, snow	logistic regression of damaged/not damaged status of National Forest Inventory plots, as assessed from aerial photos	*Picea abies*; *Pinus sylvestris*, DA	Götaland, Sweden	stand maturity; *Picea* dominance; timing of thin; admixture of *Pinus sylvestris* &/or DA trees
Albrecht et al. (2012)	wind, snow	logistic regression, classification & regression trees, of damage presence/absence & proportion of damaged b.a. in permanent sample plots	many species	Baden-Wüerttemberg, Germany	tree species; high stand height; short time from thin
Wallentin and Nilsson (2014)	wind, snow	analysis of variance of different proportions of damage between plots in a thinning trial	*Picea abies*	Halland, Sweden	high thin intensity; dec. lower stem taper

study	damage by	modelling	species or forest type	location	important predictors of inc. damage
Díaz-Yáñez, Mola-Yudego, González-Olabarria, and Pukkala (2017)	wind, snow	logistic regression of proportions of damaged, broken, or uprooted (separately), in National Forest Inventory plots	*Pinus*, *Picea* & *Betula* spp.	across Norway	large mean dbh; high mean tree height/dbh ratio; high mean height; high b.a.; species
Jalkanen and Mattila (2000)	wind, snow	logistic regression & conditional logistic regression of proportion of damage in overstorey in National Forest Inventory plots	*Pinus sylvestris*, *Picea abies* & *Betula* spp.	across northern Finland	wind: large mean dbh; high stand age; seed-tree & other dispersed cutting snow: decreasing temp. sums; high elevation; conifer dominance; mineral soils; poor drainage; high sph; pole-stage stand; close to stand edge

Table 1-2: empirical studies of wind and snow damage in the literature: modelling with spatial variables

study	damage by	modelling	species or forest type	location	important predictors of inc. damage
Wright and Quine (1993)	wind, snow	Conchran-Mantel-Haenszel tests of damage severity category per stand	*Picea*, *Pinus*, & *Larix* spp., DA	North Yorkshire, United Kingdom	high tree height; species; steep slopes
Valinger and Pettersson (1996)	wind, snow	analysis of covariance of different proportions of damage between plots in an thinning and fertilisation trial	*Picea abies*	across southern Sweden	wind: lower b.a. post-thinning; lower stand age snow: light/no thin; high latitude; high elevation; high stand age; high site index both: treatment blocks
Valinger and Fridman (1999)	wind, snow	logistic regression of proportion of damage in National Forest Inventory plots	*Pinus sylvestris*, *Picea abies* & *Betula* spp.	across all of Sweden	*Pinus sylvestris:* high elevation; lat/long; high stand age. *Picea abies:* high elevation; lat/long; high plot total volume
Mitchell, Hailemariam, and Kulis (2001)	wind	logistic regression of damage presence/absence in 50m segments of cutblock boundary, as assessed from aerial photos	*Tsuga heterophylla*, *Thuja plicata* & *Abies amabilis*	Vancouver Island, Canada	high growth potential; high sph; cutface orientation relative to prevailing wind; long time since harvest; high topographic exposure
Lindemann and Baker (2002)	wind	classification & regression trees, logistic regression, both of categorised damage percentages, of sample plots imposed on aerial photos at a 0.5 km grid	*Picea engelmanii*, *Abies lasiocarpa* & *Pinus contorta*	Colorado, USA,	inc. distance to mt. range; high wind exposure; high elevation; east aspect; *Picea* dominance.
Lanquaye-Opoku and Mitchell (2005)	wind	logistic regression of proportion damaged in 25x25 m plots along cutblock boundaries, assessed from aerial photos	*Abies amabilis*, *Thuja plicata*, *Cupressus* spp. DA, *Pseudotsuga menziesii*, *Tsuga mertensiana*, *Pinus contorta*, *Picea* spp.	British Colombia, Canada	all sites: high mean annual wind speed; W versus E aspects some sites: high wind exposure; dec. shelter; steep slopes; high tree height; high site index
Stueve, Lafon, and Isaacs (2007)	ice storm	logistic regression & chi-squared tests of categorised before-&-after-storm declines in NDVI per pixel, at 600 m spacing, from Landsat 5 images	*Quercus*-dominated forests	Appalachian Mountains, Virginia, USA	high elevation; aspect (E through S); steep slopes
Aszalós et al. (2012)	ice storm	logistic regressions of presence/absence of ice damage in a) a map compiled from field survey, & b) co-located aerial photographs, at random points spread > 40m apart	*Quercus petraea*/*Q. cerris*/*Carpinus betulus* mixtures; *Fagus sylvatica* dominant	Börzsöny Mountains, Hungary	high elevation; slope (variable outcomes); aspect (nearness to SE); high % of *Fagus sylvatica*; inc. height dominant species

Krejci, Kolejka, Vozenilek, and Machar (2018)	wind	logistic regression of presence/absence of wind damage at points on a 25x25m grid overlaid on aerial photographs	*Picea abies*	Šumava National Park, Czech Republic	high elevation; high stand age; deeper soil; high *Picea* %; wind direction; high wind speed
Díaz-Yáñez, Mola-Yudego, and González-Olabarria (2019)	wind, snow	boosted regression tree predictions of wind & snow damage as recorded in National Forest Inventory plots	Many species, predominantly *Pinus*, *Picea* & *Betula* spp.	forests across Norway	high lat.; high elevation; steep slopes; high sph; large mean dbh, high stand dominant height

Table 1-3: empirical studies of wind and snow damage in the literature: modelling with spatial equations

study	damage by	modelling	species or forest type	location	important predictors of inc. damage
Martín-Alcón, González-Olabarria, and Coll (2010)	wind	linear regression of proportion of damage, with an auto-covariate for storm regime, from National Forest Inventory plots	*Pinus nigra*, *P. sylvestris* & *P. uncinata*	Pyrenees Mountains, Spain	high topographic exposure; interaction of b.a. & height/dbh ratio
Hanewinkel, Breidenbach, Neeff, and Kublin (2008)	wind, snow	Plots from National German Forest Inventory probability of damage: logistic regression amount of damage: linear mixed regression with autoregression	Many species/forest types studied	Black Forest, Germany	high elevation; inc. soil wetness; high stand volume
Schmidt, Hanewinkel, Kändler, Kublin, and Kohnle (2010)	wind	generalised additive models with spatial trend function, analysing proportion of damage in plots from National German Forest Inventory	Groups: *Fagus* spp. & *Quercus* spp.; other DA; *Picea abies*; *Pinus sylvestris* & *Larix* spp.; *Abies alba* & *Pseudotsuga menziesii*	Baden-Wüerttemberg, Germany	high tree height; high height/dbh ratio; species (especially DA v. conifer); aspects W & SW, inc. soil waterlogging
Hanewinkel, Kuhn, Bugmann, Lanz, and Brang (2014)	wind and snow	logistic regression with an auto-covariate for presence/absence of damage, in 648 fully-censused forest stands measured every 5 – 10 years since 1920	*Picea abies*, *Abies alba*, *Fagus sylvatica* in uneven-aged forests	Neuchâtel, Switzerland	stand dbh distribution; high intensity of harvesting in 8 years before storm; high topographic exposure; slope (variable effects)

1.4.4 Explanatory variables in empirical wind and snow damage studies

1.4.4.1 Overview of potential explanatory variables

Schindler et al. (2012), writing their editorial 'Wind effects on trees', lists metrological conditions, site conditions, topographic conditions and tree and stand characteristics as influencing the probability of storm damage in forests, noting also that site and topographic features are essentially static conditions with respect to storm damage. As the detailed weather conditions during storms are often not known, as they are not in this study, many authors attempting to review and summarise the field concentrate on other correlating factors. For instance, Martin and Ogden (2006), in their review paper of wind damage in New Zealand forests, found that the main abiotic factors that influenced damage patterns were topography, soil conditions, and the history of disturbance; and the main biotic factors were tree height, tree health, position of the tree within the stand and species. Whether trees broke or uprooted was controlled by rooting depth and canopy position. In their estimation, the New Zealand findings largely agreed with studies from other countries. This is reinforced by the findings in a review article by Everham and Nicholas (1996), who note that 'the spatial pattern of damage is influenced by both biotic and abiotic factors. Biotic factors that influence severity of damage include stem size, species, stand conditions (canopy structure, density), and the presence of pathogens. Abiotic factors that influence severity of damage include the intensity of the wind, previous disturbance, topography, and soil characteristics.'

This section groups the findings of studies listed in Table 1-1 by explanatory variable type. Where useful, it also introduces on findings from review articles and meta-analyses pertaining to the interaction of wind and trees, and mechanistic studies of wind and trees. This discussion pertains entirely to measures of tree damage frequency. The sole investigation of the *height* at which trees are damaged is that of Wrathall (1989), summarised in Table 1-1, above.

1.4.4.2 Weather conditions

Given that the focus of this study is forest damage assumed to have been caused by wind or wind plus snow, wind speed at the time of tree damage would seem an obvious candidate for an explanatory variable. However, previous research has found that choosing a wind speed at which damage is likely for any given forest is not straightforward, due to the complex interactions between the wind and the forest (Quine, 1995; Ruel, 1995; Zeng, Garcia-Gonzalo, Peltola, & Kellomäki, 2010). Findings from mechanistic studies show that critical wind speed varies with the age of the stand (Moore & Quine, 2000), and different trees within the same stand have different experiences of wind, modified by their position and neighbours (Peltola, Väisänen, Kellomäki, & Ikonen, 1999). Simulations have shown that proximity to gaps is also influential (Zeng et al., 2010). A further issue arises because wind gusts or other short-run measures of wind, not available for this study, have been found to be more important to tree breakage than average wind speeds, for example by Ruel (1995) and Usbeck et al. (2012).

These issues notwithstanding, a literature overview of wind speed thresholds for tree damage was conducted, drawing on research into conifers in single-species stands or few-species stands across temperate and boreal situations: research that is specifically about radiata pine and Douglas-fir is scanty. Likewise, the use of wind speed in empirical research is rare, although Lanquaye-Opoku and Mitchell (2005) used mean annual wind speed in empirical modelling, and Krejci et al. (2018) used wind direction and wind speed in empirical modelling. The literature identifying wind speeds at which trees break arises mostly from the use of deterministic models such as ForestGALES or WINDA, or

from tree-pulling tests. Also, most of the studies deal with gust/short run average wind speeds, not wind speeds averaged over longer periods.

Potentially relevant figures located include 72 km per hour in *Abies balsamea* from tree-pulling tests (Achim, Ruel, Gardiner, Laflamme, & Meunier, 2005), 90 km per hr from the WINDA model in *Picea abies*-dominated stands (Blennow & Olofsson, 2008), 72 to 126 km per hour from the ForestGALES model in *Pinus pinaster* (Cucchi et al., 2005), 34 to 47 km per hour from tree-pulling tests in *Picea mariana* and *Pinus banksiana* (Elie & Ruel, 2005), and 106.2 km per hour decreasing with age to 65.2 km per hour from the ForestGALES model in *Pinus radiata* (Moore & Quine, 2000). Martin and Ogden (2006) in their review paper give 110 km/hr as an all-round figure for damage to occur in *Pinus radiata*, derived from historic comparisons of wind damage versus recorded wind speeds; the measurement intervals are not given, but are stated not to be gusts.

None of this rather disparate mix of figures proved very comparable to the wind speed data available for this study, which are the wind speeds at 9 am at the Timaru Aerodrome, a figure which can at best identify entire days of stronger winds. Wind speeds in this study were ultimately given a different treatment, which is detailed in *Methods*.

A possible use of precipitation records, along with air temperatures, is as a proxy for snowfall. This usage is not mentioned in the literature surveyed. Few authors even mention precipitation as a potential explanatory variable. This may be because many of the reviewed studies of wind and snow damage to trees seek to quantify damage after a single event in a relatively compact area, a scenario which has a single and fixed measure of precipitation. Exceptions include Jalkanen and Mattila (2000), whose study of the susceptibility of forest stands to wind and snow damage in northern Finland explicitly named wind and snow as non-quantified factors that were compensated for the use of matched variables in conditional logistic regression models; and Päätalo (2000), whose study of the risk of snow damage to three forest types in Finland calculated the likely temporal frequency of critical snow loads from temperature and precipitation records. In a similar fashion, barometric air pressure does not appear as an explanatory variable in the literature surveyed.

1.4.4.3 Site conditions

Amongst the findings of the studies cited in Table 1-1, the most common site conditions are measures of poor drainage or increased soil wetness, which generally increase the rate or proportion of wind damage to trees, for example Jalkanen and Mattila (2000); Hanewinkel et al. (2008) and Schmidt et al. (2010). Martin and Ogden (2006) and Mitchell (2013) in their New Zealand and worldwide review articles, respectively, propose that high soil wetness may be generally linked to the risk of wind damage to forests, especially to windthrow.

1.4.4.4 Topographic conditions

Topographic conditions are relevant as an explanatory variable in research regarding wind damage to trees largely because of their interaction with and modification of regional-scale damaging winds, as noted by Mitchell (2013). Additionally, topographic exposure and shelter control the distribution of snow over the landscape (Chapman, 2000). The importance of topography to wind effects on trees was recognised long ago. For example, Everham and Nicholas (1996) in their review article 'Forest Damage and Recovery from Catastrophic Wind', which covered studies back to 1831, found that aspect, ridges, valleys, exposed slopes, lee slopes, slope steepness, and gradient change have all been meaningful in research.

When attempting to relate wind speed to topography and geography, Hannah, Palutikof, and Quine (1995) defined these variables as relating to wind speed: elevation, topographic exposure, height of the surface elements, geographic location, and distance to the coast. In this study, the first three elements have been considered, but not geographic location and distance to the coast, which may be approximated as having a single value each for Geraldine Forest.

The most common influential topographic condition amongst the relevant literature is elevation, where increasing elevation predicts increasing damage in Valinger and Pettersson (1996), Valinger and Fridman (1999), Jalkanen and Mattila (2000), Dobbertin (2002), Lindemann and Baker (2002), Stueve et al. (2007), Hanewinkel et al. (2008), Aszalós et al. (2012), Krejci et al. (2018), and Díaz-Yáñez et al. (2019). The next most frequent variable is slope, where Dobbertin (2002), Stueve et al. (2007), and Díaz-Yáñez et al. (2019) found increasing damage with increasing slope steepness; Wright and Quine (1993) and Lanquaye-Opoku and Mitchell (2005) found decreasing damage with increasing slope steepness; and Aszalós et al. (2012) and Hanewinkel et al. (2014) found variable associations of slope with damage.

Aspect appears as frequently as slope, with Dobbertin (2002), Lindemann and Baker (2002), Lanquaye-Opoku and Mitchell (2005), Stueve et al. (2007) and Schmidt et al. (2010) uncovering significant relationships with damage, the nature of which varied, presumably because the effect of aspect depends on the wind direction at the time of damage. Equally frequent are measures of exposure, in studies by Mitchell et al. (2001), Lindemann and Baker (2002), Lanquaye-Opoku and Mitchell (2005), Scott and Mitchell (2005), Martín-Alcón et al. (2010) and Hanewinkel et al. (2014), which variously propose links between increased damaged and decreased shelter or protection, or increased damage and increased exposure (wind or topographic), or increased damaged and wind run/wind fetch.

A possible influence on tree damage at Geraldine Forest, and which is closely related to the influence of aspect, is the influence of lee slopes. Port Blakely staff suspect that the accumulation of snow on lee slopes influences damage patterns at Geraldine Forest. Martin and Ogden (2006), in their New Zealand review paper, list lee slopes as the topography most likely to incur increased wind damage in New Zealand plantation, *Nothofagus*, and *Nothofagus/*podocarp mixed forests as a finding from previous observations of wind damage.

1.4.4.5 Tree and stand characteristics

Martin and Ogden (2006) list tree height, tree health, position of the tree within the stand and tree species as important influences on wind damage in New Zealand forests. Similarly, in their worldwide review article, Everham and Nicholas (1996) list pathogens, stem size, age distribution, species mix, maturity, stocking relative to thinning history, and wind event juxtaposition with disturbances as important factors. These lists, however, are too generalised for the particular case of Geraldine Forest, which is a plantation forest of even-aged, single-species stands, without significant tree health issues.

Reviewing the studies outlined in section 1.4.2, the first most common stand characteristic that is influential for wind damage is increasing stand or tree height and/or increasing stand or tree age, where increasing heights and ages are accompanied by an increase in damage severity. Wright and Quine (1993), Veblen et al. (2001), Aubrey et al. (2007), Schmidt et al. (2010), Valinger and Fridman (2011), Aszalós et al. (2012), and Krejci et al. (2018) draw this conclusion from studies of single storm events. Valinger and Pettersson (1996), Fridman et al. (1998), Valinger and Fridman (1999), Jalkanen and Mattila (2000), Lanquaye-Opoku and Mitchell (2005), Albrecht et al. (2012), Díaz-Yáñez et al. (2017) and Díaz-Yáñez et al. (2019) draw similar conclusions from studies of permanent sample plot data, or similar, where the date(s) of damaging storms are not known. The studies after single storm

events illustrate that that increased height and age have effects in their own right, and do not only represent older areas of forest accumulating more damage as time passes.

The next most common stand characteristic that appears influential is species or species mixture, with Lindemann and Baker (2002), Schmidt et al. (2010); Wang and Xu (2009), Aszalós et al. (2012), Díaz-Yáñez et al. (2017), and Krejci et al. (2018) discovering significant differences between species or between mixtures with one dominant species. Similar findings are made by Wright and Quine (1993), Jalkanen and Mattila (2000), Veblen et al. (2001), Valinger and Fridman (2011) and Albrecht et al. (2012), and these authors also discover a difference between coniferous and broad-leaved angiosperm species, which they attribute to the reduced snow and wind effects on trees that are leafless in the winter, when damaging weather occurs.

The third most frequent finding is that increased biomass, variously expressed as increased basal area (Díaz-Yáñez et al., 2017; Martín-Alcón et al., 2010), high stocking (Fridman et al. (1998), high volume (Díaz-Yáñez et al., 2019), high site index (Lanquaye-Opoku & Mitchell, 2005; Valinger & Pettersson, 1996), or increased crown biomass (Aubrey et al., 2007; Wrathall, 1989)exhibited a higher degree of damage.

The fourth most common finding is the influence of tree shape. Munishi and Chamshama (1994), Scott and Mitchell (2005), Martín-Alcón et al. (2010), Schmidt et al. (2010), and Albrecht et al. (2012) found that increasing height to diameter ratio was predictive of increased damage, with Aubrey et al. (2007) finding it predictive of decreased damage.

The silvicultural history of a stand of trees has often been noted as interacting with wind effects. For example, Ruel (1995) gives a thorough review of the effects of silviculture on windthrow in the eastern Canadian context. The link between storm damage and timing of disturbance by thinning (or single-tree harvest) is perhaps not as commonly found in studies as reviews of the field might suggest, but Valinger and Pettersson (1996), Valinger and Fridman (2011), Albrecht et al. (2012) and Hanewinkel et al. (2014) do detail such an effect.

Sometimes the stand variables dominate the response to wind, in comparison to other variables. Albrecht et al. (2012) found, in a study of forests in south-western Germany, that storm damage could be modelled by tree type, tree age, and stand disturbance from silviculture and selection harvesting history, and soil, and site conditions and topographic variables were not influential. Díaz-Yáñez et al. (2017) had similar findings in Norway. Similar findings in this research would not be surprising, as it deals with scattered wind-damage, which Veblen et al. (2001) suggest will be more influenced by stand factors.

1.5 Core assumptions for this thesis

This research assumes the studied damage to trees at Geraldine Forest *is* in fact caused by wind or wind plus snow, as suspected by the forest managers. This assumption has strongly guided the choice of literature that has been reviewed, which has in turn guided the choices of explanatory variable used in this research. It is possible that some of the damage to trees at Geraldine Forest is caused by other factors. Efforts to model such damage using variables relevant to wind and snow are unlikely to succeed.

The models developed in this thesis are empirical models, not process models. In other words, the models use variables that correlate with damage. Empirical models have limitations in that they do not examine casual links, and should not be applied at sites with different biophysical or management

characteristics to their source data (Mitchell et al., 2001). However, proving the link between wind (or wind plus snow) and tree breakage would require controlled experiments, which are rare in this field; three studies listed in section 1.4.3 are exceptions, namely Valinger and Pettersson (1996), Aubrey et al. (2007), and Wallentin and Nilsson (2014).

This thesis does not deal with large-area wind damage at Geraldine Forest, where the majority of trees are broken or uprooted across a substantial area. As damage frequencies increase, at some point an area ceases to be an area of standing trees with scattered wind damage and becomes an area of wind damage with scattered standing trees. When this point is reached by the estimation of Port Blakely, the affected area is mapped out, the stand area is written down, and the affected area is either salvage harvested, cleared and re-planted, or left until the surrounding area is felled and then prepared for replanting along with the surrounding stand. The tree data underpinning this thesis are from ground inventory plots that fell in stands, not in areas that had been mapped out due to wind damage. Port Blakely undertake regular remapping from aerial photos and local knowledge to exclude such areas. Therefore these research data are from areas that were expected, at the time of inventory, to be stands, and these findings apply to stands.

This study uses the term weather, rather than climate. Geraldine Forest experiences approximately the same climate in all areas, perhaps with some variation imposed by elevation differences. Different plots do, however, experience different weather, as they occupy different windows of time amongst the weather history of the forest, which in turn influences the coincidence of vulnerable periods in the plots' life history with potentially damage-causing weather.

Finally, this research undertaken largely to investigate damage to radiata pine. Damage to Douglas-fir, which is also grown at Geraldine Forest but suffers less damage, is studied and investigated partly as a comparison, and partly to indicate to forest managers the magnitude of damage that might occur, were the radiata pine crop to be replaced with Douglas-fir.

2 Methods

Having explored the research topic, the research questions, and the relevant literature in Chapter One, this chapter details the data preparation, exploratory analysis and statistical methods undertaken to create the Results presented in Chapter Three.

2.1 Study site

The overall study site for this research is Geraldine Forest, in the Canterbury region of New Zealand. Geraldine Forest, owned and managed by Port Blakely New Zealand Forestry, occupies 5,580 hectares of hill country, with its approximate centre at 171.079 E -44.081 S, and has an elevation range from 147 metres to 857 metres above mean sea level. The forest is in the Eastern South Island climate zone, which is heavily influenced by the Southern Alps to the west, having low mean annual rainfall with summer dry periods, typical summer daytime maximum temperatures of 18°C to 26°C, typical winter maximum air temperatures from 7°C to 14°C, and winter frosts (National Institute of Water and Atmospheric Research, 2019a). The 20-year average climate data for a virtual climate station (National Institute of Water and Atmospheric Research, 2019b) within the forest are shown in Table 2-1, below. A more extensive table with minima, maxima and standard deviations is available in Appendix 6.4.5.

Table 2-1: Weather data 01/01/1997 to 31/12/2016, for Virtual Climate Station number 15231.

Month	mean daily maximum temperature, °C	mean daily minimum temperature, °C	accumulated precipitation, mm, 20 year mean	mean daily mean windspeed, km/hr
January	21.4	9.7	74.6	9
February	21.2	9.7	61.6	6.5
March	19.6	7.7	53.7	7.1
April	16.3	4.8	68.7	6.5
May	13.4	2.7	60	6.9
June	10.6	-0.4	50.8	6.5
July	10.1	-0.9	50.6	7.1
August	11.4	0.5	65.1	7.4
September	14.4	2.6	not available - source data error	8.0
October	16.2	4.3	66.5	9.8
November	17.9	6.1	64.2	7.6
December	19.9	8.6	69.9	8.2

Geraldine Forest is planted in a mixture of species. The most common species is the exotic conifer *Pinus radiata* (radiata pine), which is native to coastal California, U.S.A, and islands off the Baja California Peninsula, Mexico. The second most common species is the exotic conifer *Pseudotsuga menziesii* (Douglas-fir), which is native to the Pacific North-west of Canada and the United States. There were approximately 2300 hectares of radiata pine and 1400 hectares of Douglas-fir in the forest at the end of the study period (31/12/2016).

There were three different original data sources: inventory plots, LiDAR ground control plots and permanent sample plots. All three types are circular, bounded plots of known location. Inventory plots are non-repeated measurement plots from conventional ground-based mid-rotation or pre-harvest inventory. Their original purpose was to provide data on stands at the time of inventory, in a format and at a sampling rate that could be used to model the growth of the stand forward for yield estimation purposes. LiDAR ground control plots are similar to inventory plots, but instead of being

directly modelled for yield, they are related to tree metrics calculated from a LiDAR point cloud, and the two are modelled forward together. Permanent sample plots are repeated measurement plots; for this study, to avoid pseudoreplication from permanent sample plot data, a single plot measurement was chosen that is a) after completion of all plot silviculture and b) has a complete description of the tops of trees. The numbers of plots by type are shown in Table 2-2, below. Inventory plots are the dominant data source, and the only data source for Douglas-fir.

Table 2-2: plots by source data type and species.

	number of plots by source data type		
species	inventory	LiDAR	permanent sample plot
radiata pine	418	198	9
Douglas-fir	317	0	0

The study sites are distributed across the forest, as shown in Figure 2-1, below.

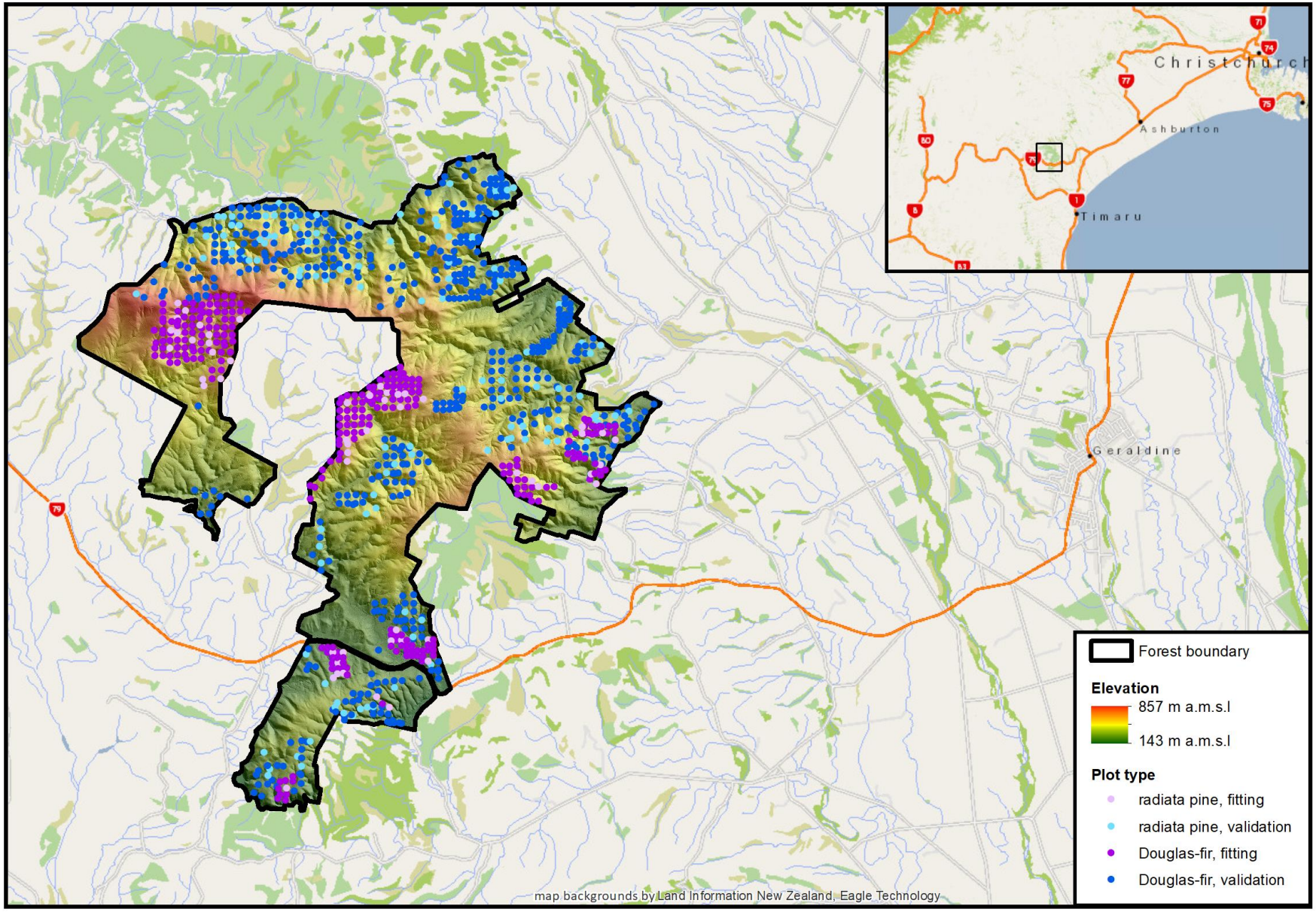

Figure 2-1 Geraldine Forest: locality map and plot distribution map, showing model fitting and validation plots by species.

2.2 Methodological overview

First, the following data were gathered:

1. The locations of all forest measurement plots were created as points in an ArcGIS geodatabase feature class.
2. Tree description variables, for example tree height, were assembled from the forest measurement plots, first at an individual tree level, then aggregated to plot-level variables.
3. Potential topographic variables, for example aspect, were calculated at the plot level.
4. Variables describing establishment (planting and associated operations) and silviculture were assembled at the plot level; their values are identical for plots that are located within the same stand, i.e. areas with a shared establishment and silvicultural history and general location.
5. Weather variables were calculated at the plot level, from a single climate station representing the study area. These variables have differing values because the timespan from planting to measurement varies by plot.

Second, visualisation and exploratory data analysis was undertaken to identify the potentially important variables relating to standing tree breakage and windthrows.

Third, the mean levels of damage in radiata pine and in Douglas-fir were tested for statistically significant differences.

Fourth, the plots were split into a fitting data set (85%) and a validation data set (15%).

Fifth, random forests and regression analyses were used on the fitting data to create models of the response variables plot mean broken height (*P_BRKN_ht_mean*), proportion of trees damaged per plot (*Tops_prpn_DAM*), and proportion of live trees per plot (*Prpn_LIVE*). These models were then tested against the validation data.

Data compilation for this thesis was conducted in a mixture of Excel (Microsoft Office, Redmond, Washington), R version 3.6.1 (R Core Team, 2019) running in R Studio version 1.2.1335, and ArcGIS for Desktop (ESRI, 2017). Exploratory data analysis and graphing, and statistical analysis and graphing, were undertaken in R.

2.3 Sources of raw data

The primary digital terrain model, the forest measurement plot data, and the stand records used in this study were sourced from Port Blakely, the supporting company, under a data use agreement to access Port Blakely's private data. Weather data were sourced from the National Institute of Water and Atmospheric Research and the New Zealand Metrological Service, New Zealand government entities that hold weather data archives. The secondary digital terrain model is web-hosted by the University of Otago School of Surveying and is available for public download. Table 2-3, below, gives further detail.

All data required considerable compilation and manipulation to become a set of variables suitable for analysis. A brief description of this process is given in section 2.4.2, with more detailed information available in the Appendices.

Table 2-3: data sources for this study.

data	data generation process	data scope	data provider	Data status
primary digital elevation model, one-metre resolution	calculated from LiDAR point cloud, single-swath mean density 19.52 points per m^2	continuous across Geraldine Forest	Port Blakely	private
NZSoSDEM v1.0 (secondary digital elevation model) 15-metre resolution	interpolated from Land Information New Zealand 20m photogrammetric contours	continuous across New Zealand	University of Otago School of Surveying	public
forest measurement plots	detailed numeric descriptions of trees within bounded circular plots; plots have known measurement date and location	conventional mid-rotation and pre-harvest inventory plots, 426 in radiata pine and 317 in Douglas-fir 190 plots from LiDAR survey ground control, in radiata pine nine permanent sample plots, in radiata pine	Port Blakely	private
stand records	planting, pruning and thinning recorded in stand record system	for each forest stand	Port Blakely	private
weather data	daily weather records, Timaru Aerodrome	single location	New Zealand Metrological Service	public, access charge
weather data	daily weather data interpolations, Virtual Climate Station number 15231	single location	New Zealand Institute of Water and Atmospheric Research	public, access charge

2.4 Variables used in this research

2.4.1 Choice of variables

The choice of variables used in this research was driven partly by data availability and partly by suggestions from the literature, especially the list given in Martin and Ogden (2006), in their review paper of wind damage in New Zealand forests. Table 2-6, below, compares that list and the variables included in this research.

Table 2-4: Comparison of variables, Martin and Ogden (2006) and this study.

Martin and Ogden (2006)		**this study**
abiotic variables	topography	various included
	soil conditions	not included (no data)
	history of disturbance	silvicultural history regarding thinning and pruning tree variables regarding forking and live/dead status
biotic variables	tree height	included in tree description variables
	tree health	not explicitly included, tree live/dead status may relate
	canopy rank of the tree	not included (trees studied are co-dominants)
	species	included

2.4.2 Nature of the variables

2.4.2.1 Plot description variables

The variables listed in Table 2-5, below, describe the plots, either individually (*Plot_no*), or as a group by the stand identity (*P_stand*), the planting year (*P_YOE*), or the measurement year (*P_YOM*) to which they belong. Any group by *P_stand* will have a single *P_YOE* and *P_YOM*, but the reverse does not apply. See section 2.6.2 for a description of the hierarchy among the variables used in this study. Plot size is used only to calculate plot stocking and basal area per-hectare equivalents (see Table 2-6).

Table 2-5: variables used to describe plots.

variable	**definition**
Plot_no	Plot number. A standardised version of the original plot number assigned at plot survey time. Alphanumeric. Not visualised in EDA.
P_stand	The identity of the stand within which the plot occurred. Alphanumeric.
P_YOM	Plot year of measurement. Factor with eight levels. Not to be confused with P_meas_date.
P_YOE	Year of establishment (planting) of the trees in the plot. Factor with 28 levels.
P_size	Plot size, in hectares. Numeric.

2.4.2.2 Tree description variables

Tree description variables, as described in Table 2-6, were calculated at the plot level from measurements of individual trees. Data from PSPs contains no information about forks. Therefore, the values for forking variables for PSPs were entered as NA (missing data).

Calculating tree age at measurement depends upon knowing their planting date. In this study, most plots have only their planting year recorded, and so all planting dates are set to July 1 of the relevant year. This is the standard assumption in New Zealand forestry when the planting year, but not date, is known. Radiata pine and Douglas-fir are generally planted during their winter dormancy. Under Geraldine Forest's local conditions, this could range from 1 June to 1 September.

Mean measured tree heights were calculated only for the tree top statuses 'NRML' (normal top) and 'BRKN' (broken top). The category 'HORZ' (horizontal top) means that the tree is lying on its side because of windthrow and therefore does not have a meaningful height.

Some assumptions were made about pruned heights. The LiDAR and conventional inventory data captured the presence of 'epicormics' – small branches that have re-grown in the middle of a pruned section of stem, from a bud associated with the pruned branch stub. These branches occurred at variable heights, giving some trees with two pruned sections, namely above and below the epicormics. There were also some instances where the tree forks and both forks had been pruned. The bearing that pruning has on this study relates to the wind dynamics of a pruned (bare) stem. Therefore, some decision rules were developed about what to record for the pruned height. These were:

- An epicormic branch in the 1 cm class is unlikely to much affect the wind dynamics of the tree, and so the pruned height is the higher of the two pruned heights for the tree.
- An epicormic branch in the 4 cm or 7 cm (or larger) classes could begin to affect the wind dynamics of the tree, and so the pruned height is the lower of the two pruned heights for the tree.
- Epicormics at 0.3 m and below were ignored, as they are practically at ground level and will have minimal interaction with wind, and the pruned height was whatever was given in the tree description.
- The pruned height of a forked tree with both forks pruned is the lower of the two pruned heights.

Tops_prpn_DAM is a composite variable, including trees with the top status *BRKN* (broken) or *HORZ* (tree windthrown and lying on its side). Of the two, broken tops are far more common. In radiata pine, considering only trees whose tops were assessed, the top codes are BRKN: 2910 trees, HORZ: 263 trees, and NRML (normal top): 3795 trees. For Douglas-fir, the figures are BRKN: 385 trees, HORZ: 28 trees and NRML: 1038 trees. Therefore, broken tops are much more common than windthrow in attritional damage at Geraldine Forest.

Table 2-6: variables used to describe trees.

variable	definition
P_BRKN_ht_mean	Mean height of broken trees, in metres. *Also a response variable.*
Tops_prpn_DAM	Proportion of trees with tops damaged (BRKN plus HORZ). *Also a response variable.*
Prpn_LIVE	Proportion of live trees. *Also a response variable.*
P_count	Number of trees in the plot.
P_date_meas	Date at which the plot was measured.
P_age_meas	Age at which the plot was measured, in years.
P_sp	Plot tree species. One per plot.
P_dbh_mean_NRML	The arithmetic mean of the diameters at breast height (1.4 m) of the normal-top trees in the plot, in millimetres.
P_dbh_mean_ BRKN	The arithmetic mean of the diameters at breast height (1.4 m) of the broken-top trees in the plot, in millimetres.
P_tree_ht_mean_NRML	The arithmetic mean of the height of the normal-top trees in the plot, in metres.
P_tree_ht_mean_BRKN	The arithmetic mean of the height of the broken-top trees in the plot, in metres.
P_BA_ha_equiv	The basal area of the trees in the plot, converted to a per-hectare equivalent (necessary because plots were of various sizes).
P_sph_equiv	The stocking of trees in the plot, converted to a per-hectare equivalent (necessary because plots were of various sizes).
P_slend_mean	Arithmetic mean slenderness of the plot trees, where slenderness = mean height trees with normal tops/mean diameter trees with normal tops. Unitless ratio.
P_Fk_1_prpn	Proportion of trees that fork once in the plot.
P_Fk_1_ht	Arithmetic mean height at first fork, in metres.
P_Fk_2_prpn	Proportion of trees that fork twice in the plot.
P_Fk_2_ht	Arithmetic mean height at second fork, in metres.

2.4.2.3 Silvicultural history variables

The silvicultural history variables given in Table 2-7 were calculated at the plot level from silvicultural history information, except for P_pru_prpn and P_pru_ht, which were calculated from tree measurements. Due to incompleteness of the silvicultural history information, some assumptions were made to make the silvicultural histories as complete as possible; see Appendix 6.1.5 for details.

Table 2-7: variables used to describe silviculture.

variable	definition
P_pruned	Plot has been pruned, or not. (P/NP).
P_pru_prpn	Proportion of trees that are pruned in the plot.
P_pru_ht	Arithmetic mean pruned height of trees in the plot, in metres.
P_thinned	Plot has been thinned, or not, or unknown (T/UT/NA). Calculated as T if number of thinnings >= 1.
Age_thin	Age at thinning, in years, from silvicultural records.
Age_LastP	Age at the last pruning event, in years, from silvicultural records.
Estab_sph	Stocking at stand establishment (planting), in stems/ha, from silvicultural records.
Final_sph	Stocking after final thin, in stems/ha, from silvicultural records.
Sph_drop	Final_sph as a proportion of Estab_sph, unitless
T_P_gap	Time elapsed between Age_LastP and Age_thin, in years (is negative if thinning was completed earlier than pruning).

2.4.2.4 Topographic variables

Topographic variables created for this research are given in Table 2-8. The plot-level values for the topographic variables elevation, aspect, slope and the various morphometric protection indices were extracted for plots from continuous spatial surfaces. These surfaces have one-metre resolution, meaning that using plot point locations to extract plot-level topographic variables could be misleading, as micro-topography could return values different to the average for the plot. Therefore, a two-step process was used to create plot-level values.

First, plots were buffered to create polygons matching the plot footprint. Second, the topographic variables of plot mean elevation, plot mean slope, and plot predominant aspect were extracted for each plot footprint. Elevation (*P_alt*) and slope (*P_slope*) are simple arithmetic means of the values in each footprint.

Because aspect is a circular variable, where 0 degrees and 360 degrees are both north, aspect requires transformation before it can be used in any meaningful way. Therefore, the original numeric aspect raster was transformed into classified rasters to represent cardinal directions. Three alternative aspect classifications were created, and the mode of the classifications in each plot footprint was assigned as the aspect of that plot. For specifics of the calculations, see Appendix 6.2.3. The aspect classifications are:

- *card_4way_N*, which is divided into four to give north, east, south, and west aspects
- *card_4way_NE*, which is divided into four (at 45 degrees offset to *card_4way_N*) to give north-east, south-east, south-west, and north-west aspects
- *card_8_way*, which is divided into eight to give north, north-east, east, south-east, south, south-west, west, and north-west aspects

Because the spatial resolution of the base aspect raster for Geraldine Forest has one-square-metre pixels, each plot footprint potentially has more than one aspect underlying it. The predominant aspects assigned to plots could, in theory, be based on areas as low as 25% of the total area for the four-way cardinal aspects and 12.5% of the total plot area for the eight-way cardinal aspect. An analysis of predominant aspect was conducted, using the calculation *proportion = plot predominant aspect area / plot total area*. This analysis found that proportions were in most cases nearer 1 than the theoretical minima, and therefore mis-assigned aspects at the tree level were not a large concern; no plots were excluded on the basis of this check. See Appendix 6.2.5.2 for details of this analysis and figures illustrating the distribution of proportion values.

A check of the slope raster showed that 3.8 % of the total forest area has a slope of 5 degrees or less. Therefore, there is a near-zero occurrence of ground that has no meaningful aspect. Likewise, only eight plots, out of 942, have an average slope of 5 degrees or less. Therefore, 'flat' has not been included as a value of aspect in this study.

The literature review (see Chapter 1) indicated that variables expressing how sheltered a landscape position is sometimes have bearing on tree damage results. The morphometric protection index (MPI) (Yokoyama, Shirasawa, & Pike, 2002) was chosen as a possibly useful expression of shelter. MPI was chosen, in comparison to other related indices, because it does not require wind direction as an input, which is useful given that the direction of damaging winds at Geraldine Forest is suspected rather than known. Because the MPI requires a specified calculation horizon or radius, several radii were chosen for use as alternative inputs during analysis. MPI calculations yield a value, between 0 and 1, which expresses the positive openness of a location in comparison to the landscape at the specified radius, where zero is completely sheltered (a hollow compared to the topography at the calculation radius)

and one is completely exposed (the top of a hill that is higher than all other locations within the calculation radius).

The existing digital elevation model for Geraldine Forest stretches only a short distance beyond the forest boundary, because that was the extent of the LiDAR survey providing its base data. Therefore, it was necessary to create a composite DEM of greater geographic extent, so that plot locations close to the forest boundary could still have the MPI calculated. This was achieved by extracting a portion of the publicly-available 15 metre resolution New Zealand-wide DEM (University of Otago School of Surveying, 2011) and combining it with the Geraldine DEM, using ArcMap's Mosaic to New Raster tool to output a new raster at 1-m resolution. This involved resampling the New Zealand-wide DEM to one metre, so that areas 15 x 15 cells all held the same elevation value.

MPI was calculated from the composite DEM, at radii of 100 m, 200 m, 500 m, 1000 m, and 2000 m, in the System for Automated Geoscientific Analyses (SAGA) (Conrad et al., 2015). Raster outputs from SAGA were read back into ArcGIS for visualisation, and for extraction of the average MPI for each plot by averaging values within each plot footprint. This processes caused extreme tiling artefacts in the calculation of MPI for areas outside the forest boundary, and slight tiling artefacts inside the forest boundary. This was preferable, however, to the alternative of resampling the Geraldine DEM to 15 m inside the forest boundary. See Appendix 6.2.4 for further details of the calculation of average MPI and the tiling artefact.

Two datasets for wind exposure index were available from a third party, as a list of values by plot, where values under 1 indicate wind shadowed areas and values above 1 indicate areas exposed to wind. These data were calculated from the same DEM as used in this study, for the directions north-east (22.5 degrees to 67.5 degrees) and south (from 202.5 degrees to 157.5 degrees), at a 100 m horizon, with the value returned for the plot centre.

Table 2-8: variables used to describe the topography.

variable	description of variable
POINT_X	Plot location, easting, New Zealand Transverse Mercator coordinates, in metres.
POINT_Y	Plot location, northing, New Zealand Transverse Mercator coordinates, in metres.
P_alt	Plot average elevation, in metres.
card_4way_N	Plot aspect, classified to north (N), south (S), east (E), or west (W).
card_4way_NE	Plot aspect, classified to north-east (NE), south-east (SE), south-west (SW), or north-west (NW).
card_8way	Plot aspect, classified to north (n), north-east (ne), east (e), south-east (se), south (s), south-west (sw), west (w) or north-west (nw).
MPI_100	Morphometric protection index to a 100 m horizon, assessed from plot centre. Unitless index. A value of zero indicates complete shelter; a value of 1 indicates complete exposure.
MPI_200	As for *MPI_100*, but to a 200 m horizon.
MPI_500	As for *MPI_100*, but to a 500 m horizon.
MPI_1000	As for *MPI_100*, but to a 1000 m horizon.
MPI_2000	As for *MPI_100*, but to a 2000 m horizon.
WindSheltS1	Wind exposure with regard to south, to a 100 m horizon.
WindSheltNE1	Wind exposure with regard to north-east, to a 100 m horizon.

2.4.2.5 Weather variables

This study of Geraldine Forest is different to many other wind damage studies in that the dates of damage-causing weather events are unknown: the damage measured for a plot is the accumulation of the damage over the plot's life. Damage cannot be related to particular events, nor can the return period of damaging storms, as defined in Mitchell (2013), be calculated in this research. An assumption was made that some unknown severity threshold for weather events must be exceeded to cause

damage to trees. Therefore, the weather records for Geraldine Forest have been subject to a thresholding exercise, where the resultant explanatory variables, as listed in Table 2-9, were calculated as 'number of unfavourable weather days' experienced in a plot's life from age five to age of measurement. Single-variable thresholds choose the worst 2% of days, by calculation from the 98th or 2nd percentile, whichever is relevant. For example, the 98th percentile threshold for wind speed is 29.7 km/hr: the thresholding exercise counts the number of unfavourable weather days at or above this threshold. Combined thresholds were set at levels that yield approximately 200 measures. These unfavourable weather variables elaborate on the more straightforward variable plot age, by accounting for some plots having lived through more unfavourable weather days than others.

Weather variables were calculated at the plot level from forest-level data. Weather data for Geraldine Forest were from two sources: the metrological station at Timaru Aerodrome, some 27 km distant, and the Virtual Climate Station Network (VCSN) data (National Institute of Water and Atmospheric Research, 2019b) for a single virtual weather station (number 15231, at 1446592 E 5112209 N) within Geraldine Forest's boundary. The VCSN is a spatial weather modelling and interpolation system, calculated using data from New Zealand's current and historic official meteorological stations. It constitutes the best weather data available for New Zealand, unless a site of interest happens to be very near a long-term full-coverage meteorological station, which Geraldine Forest is not. This study assumes that that Geraldine Forest is small enough so that all of the forest experiences any given weather event occurring in the region, without differentiation. This assumption has been made for simplicity of data acquisition and simplicity of modelling.

The basic data extracted from the VCSN were daily measures of maximum temperature, minimum temperature, accumulated precipitation over 24 hours, and the 9 am barometric air pressure, over the period 01/01/1972 to 25/11/2018. As the VCSN wind speed data date back only to 01/01/1997; the wind speed data for Timaru Aerodrome were used as the next best alternative. The 9 am average[2] wind speed at Timaru was extracted for the period 01/01/1970 to 31/12/2016 (the last plot measurement included in this study was in August 2016).

Weather data are available from 31/12/1971, but the planted date of the trees ranges back to 01/07/1962. For calculation of variables that express how much adverse weather a stand has experienced, the variable must apply to a consistent amount of the stand's life. Ideally this would be planting until measurement, but using planting date in this manner removes 75 plots from the input data, which is an unacceptably high loss, especially as these are all the plots for five stands, and thus remove the coverage of the data set from some geographic areas. Instead, variables expressing adverse weather events were calculated from age 5. This reduces the plots lost to modelling to 25, in two stands. This choice of age 5 follows Somerville (1995), who considered that wind damage to stands under 5 years old would be largely in the form of leaning stems, not breakage or windthrow. This accords with a review of the permanent sample plot data for Geraldine Forest, the only available sequential measures of the *same* trees, that shows that trees begin to be classified as having top damage around age 7 – 10 years.

None of the suggested values for damaging windspeeds available from the literature (see section 1.4.4.2) were directly comparable to the data available for Geraldine Forest. On the whole, the averages from the 9 am Timaru Aerodrome dataset are much lower than figures given in the literature. It seems likely that the fixed timing of the measurement interval does not capture high winds.

[2] Averaged over either 10 minutes or one hour, depending on the various practices applied at the time

Therefore a variable *u_wind_tim* was created, which counts on a per-plot basis days of with the top 2% of recorded daily 9 am wind speeds, which equates to speeds at or above 29.7 km/hr.

As wet soils can predispose trees to wind damage (see section 1.4.4.3), a variable *u_rain* was created, which identifies days of high daily accumulated precipitation. High precipitation serves as a proxy for wet soils, for which there are no direct data for Geraldine Forest. This variable counts, on a per-plot basis, days of with the top 2% of recorded daily rainfall accumulations, which equates to daily accumulations at or above 24.2 mm.

To identify especially cold days, a variable *u_min_temp* was created, which counts on a per-plot basis days of with the bottom 2% of estimated daily temperatures, which equates to temperatures at or below - 3.9 ^{o}C.

As snow is suspected to be a cause of tree damage at Geraldine Forest (see section 1.4.4.2), and involves precipitation as well as cold temperatures, a variable *u_mint_rain* was created with the intention of it being a proxy for snowfall, with threshold values of 3^{o}C for minimum daily temperature and 10 mm/day for precipitation, to identify 238 days falling above the this combined threshold, which were then further counted on a per-plot basis. 3^{o}C was chosen because the forest has a strong elevation range, and the weather data were derived from a single VCSN station at approximately 220 m altitude, a minimum above zero was set to allow for the probable occurrence of lower minimum temperatures and therefore snow at higher altitude. The 10 mm (about 1.25 standard deviations above the mean) was chosen to accumulate approximately 200 values.

Low barometric air pressures are usually associated with bad weather, although the use of barometric air pressure as a potential explanatory variable does not appear in the literature examined. A variable *u_air_pr* was created, which counts on a per-plot basis days of with the bottom 2% of barometric pressure, which equates to 9 am barometric pressures at or below 990 hPa.

A variable *u_mint_rain*, which identifies days of high wind and high daily accumulated precipitation was created, with threshold values of 14 km/hr for windspeed and 10 mm/day for precipitation, to identify 206 days falling above the this combined threshold, which were then further counted on a per-plot basis.

Table 2-9: variables used to describe the weather.

variable	description of variable
u_wind_tim	Count of highly windy days experienced during the plot's life from age 5 to measurement age. Top 2% of 9am windspeeds, taken from the Timaru aerodrome metrological station.
u_rain	Count of highly rainy days experienced during the plot's life from age 5 to measurement age. Top 2% of total daily precipitation, from VCSN 15231.
u_min_temp	Count of lowest minimum temperature days experienced during the plot's life from age 5 to measurement age. Bottom 2% of minimum daily temperatures, from VCSN 15231.
u_mint_rain	Count of days of both low minimum temperature (<= 0 ^{o}C) and high rainfall (>= 10 mm daily accumulation) experienced during the plot's life from age 5 to measurement age, from VCNS 15231.
u_air_pr	Count of lowest barometric pressure days experienced during the plot's life from age 5 to measurement age. Bottom 2% of minimum daily temperatures, from VCSN 15231.
u_rain_wind_tim	Count of days of both high rainfall (>= 10 mm daily accumulation, from VCS 15231) and high windspeed (> 14 km/hr 9am windspeed at Timaru aerodrome) from age 5 to measurement age.

2.5 Exploratory analysis

2.5.1 Summary statistics by variable

Table 2-10 and Table 2-11, below, present the summary statistics for variables used in this research. Please see also section 2.4 for a description of the variables. *P_tree_ht_mean_BRKN* (plot mean broken height), *Tops_prpn_DAM* (the proportion of damaged trees per plot), and *Prpn_LIVE* (the proportion of live trees per plot) are the response variables, and also act as explanatory variables for one another. The remaining variables are all explanatory.

Table 2-10: summary statistics by variable for radiata pine: 625 plots.

Continuous and count variables	***describes***	***minimum***	***mean***	***maximum***	***std. dev.***	***missing data count***
response variables						
P_tree_ht_mean_BRKN (m)	trees	2.6	14.5	28.1	4.7	0
Tops_prpn_DAM (proportion)	trees	0	0.432	1	0.229	0
Prpn_LIVE (proportion)	trees	0.063	0.984	1	0.039	0
explanatory variables						
P_size (ha)	plot	0.040	0.057	0.100	0.008	0
P_count (count)	trees	2	*(na)*	61	*(na)*	0
P_date_meas (date)	trees	9/05/2003	*(na)*	17/09/2016	*(na)*	0
P_age_meas (years)	trees	14.91	23.9	46.93	3.85	0
P_dbh_mean_NRML (mm)	trees	272	480	833	88	7
P_dbh_mean_BRKN (mm)	trees	113	400	708	89	77
P_tree_ht_mean_NRML (m)	trees	16.4	30.5	45.0	4.9	2
P_tree_ht_mean_BRKN (m)	trees	2.6	14.5	28.1	4.7	110
P_BA_ha_equiv (m^2/ha)	trees	4.8	50.8	93.8	15.5	0
P_sph_equiv (stems/ha)	trees	33	339	1325	144	0
P_slend_mean (ratio tree ht/tree dbh)	trees	0.029	0.065	0.100	0.009	7
P_Fk_1_prpn (proportion)	trees	0.00	0.16	1.00	0.13	101
P_Fk_1_ht (m)	trees	1.5	8.4	28.1	4.5	0
P_Fk_2_prpn (proportion)	trees	0	0.01	0.15	0.02	570
P_Fk_2_ht (m)	trees	3	11.3	26.8	5.5	0
P_pru_prpn (proportion)	silviculture	0	0.69	1	0.34	0
P_pru_ht (m)	silviculture	0	5.1	7.5	2.2	0
Age_thin (years)	silviculture	5.75	7.83	11.93	1.48	199
Estab_sph (stems/ha)	silviculture	331	1029	2240	225	0
Final_sph (stems/ha)	silviculture	261	383	649	83	87
Sph_drop (proportion)	silviculture	0.246	0.390	0.847	0.107	87
T_P_gap (years)	Silviculture	-0.080	0.847	3.250	1.107	0
POINT_X (coordinate)	topography	1441346	*(na)*	1451129	*(na)*	0
POINT_Y (coordinate)	topography	5110085	*(na)*	5122143	*(na)*	0
P_alt (m)	topography	157	419	812	115	0
P_slope (degrees)	topography	2.6	24.7	41.9	7.4	0

Continuous and count variables	***describes***	***minimum***	***mean***	***maximum***	***std. dev.***	***missing data count***
MPI_100 (unitless)	topography	0.033	0.190	0.417	0.067	0
MPI_200 (unitless)	topography	0.034	0.203	0.431	0.074	0
MPI_500 (unitless)	topography	0.034	0.217	0.431	0.077	0
MPI_1000 (unitless)	topography	0.034	0.225	0.436	0.078	0
MPI_2000 (unitless)	topography	0.034	0.229	0.442	0.078	0
WindSheltS1 (unitless)	topography	-0.264	0.384	0.787	0.204	0
WindSheltNE1 (unitless)	topography	-0.309	0.208	0.763	0.233	0
u_wind_tim (days)	weather	52	110	209	24	0
u_rain (days)	weather	62	129	170	26	0
u_min_temp (days)	weather	76	177	240	28	0
u_mint_rain (days)	weather	11	27	40	8	0
u_air_pr (days)	weather	81	157	203	25	0
u_rain_wind_tim (days)	weather	31	61	108	12	0
Categorical variables						
Plot_no (alphanumeric)	plot	625 levels				
P_stand (alphanumeric)	plot	67 levels				
P_YOM (categorical)	plot	8 levels				
P_YOE (categorical)	plot	20 levels				
P_sp (categorical)	plot	1 level (PRAD)				
P_pruned (categorical)	silviculture	P: 543	NP: 82			
P_thinned (categorical)	silviculture	T: 564	UT: 44			
card_4wayN (categorical)	topography	N: 264	E: 173	S: 73	W: 115	
card_4wayNE (categorical)	topography	NE: 243	SE: 89	SW: 85	NW: 208	
card_8way (categorical)	topography	n: 142	ne: 115	e: 93	se: 43	
		s: 27	sw: 48	w: 51	nw: 106	

Table 2-11: summary statistics by variable for Douglas-fir: 317 plots.

Continuous and count variables	***describes***	***minimum***	***mean***	***maximum***	***std. dev.***	***missing data count***
P_size (ha)	plot	0.03	0.04	0.05	0.01	0
P_tree_ht_mean_BRKN (m)	trees	3.8	15.4	28.3	4.4	0
Tops_prpn_DAM (proportion)	trees	0.000	0.221	0.833	0.224	0
Prpn_LIVE (proportion)	trees	0.065	0.969	1	0.05	0
P_date_meas (date)	trees	5/12/2011	*(na)*	25/11/2015	*(na)*	0
P_age_meas (years)	trees	35.36	40.02	50.44	4.49	0
P_count (count)	trees	2	*(na)*	39	*(na)*	0
P_dbh_mean_NRML (mm)	trees	220	375	544	52	2
P_dbh_mean_BRKN (mm)	trees	121	331	511	62	149
P_tree_ht_mean_NRML (m)	trees	14.2	27.4	37.4	3.6	2
P_tree_ht_mean_BRKN (m)	trees	3.8	15.4	28.3	4.4	149
P_BA_ha_equiv (m^2/ha)	trees	2.5	51.5	81.0	13.4	0
P_sph_equiv (stems/ha)	trees	50	506	900	157	0
P_slend_mean (ratio)	trees	0.052	0.075	0.105	0.01	2
P_Fk_1_prpn (proportion)	trees	0	0.034	0.5	0.06	0
P_Fk_1_ht (m)	trees	1.5	8.4	20.6	3.8	0
P_Fk_2_prpn (proportion)	trees	0	0.001	0.050	0.018	0
P_Fk_2_ht (m)	trees	0	0	0	0	0
Age_thin (years)	silviculture	16.18	16.45	17.43	0.50	256
Estab_sph (stems/ha)	silviculture	1250	2043	2990	599	31
Final_sph (stems/ha)	silviculture	593	648	659	22	215
Sph_drop (ratio)	silviculture	0	0.332	0.474	0.157	193
POINT_X (coordinate)	topography	1441685	*(na)*	1450400	*(na)*	0
POINT_Y (coordinate)	topography	5110002	*(na)*	5119538	*(na)*	0
P_alt (m)	topography	191.2	543.5	503.8	811.6	0
P_slope (degrees)	topography	3.3	23.6	22.7	39.8	0
MPI_100 (unitless)	topography	0.035	0.178	0.183	0.457	0
MPI_200 (unitless)	topography	0.038	0.191	0.195	0.466	0
MPI_500 (unitless)	topography	0.038	0.209	0.208	0.473	0
MPI_1000 (unitless)	topography	0.038	0.213	0.213	0.472	0
MPI_2000 (unitless)	topography	0.045	0.215	0.216	0.473	0
WindSheltS1 (unitless)	topography	-0.322	0.217	0.212	0.729	0
WindSheltNE1 (unitless)	topography	-0.253	0.327	0.328	0.835	0
u_wind_tim (days)	weather	216	264	274	381	0
u_rain (days)	weather	223	239	251	319	0
u_min_temp (days)	weather	266	281	289	359	0
u_mint_rain (days)	weather	50	53	55	69	0
u_air_pr (days)	weather	229	256	258	315	0
u_rain_wind_tim (days)	weather	120	142	149	202	0

Categorical variables				
Plot_no (alphanumeric)	plot	317 levels		
P_stand (alphanumeric)	plot	18 levels		
P_YOM (categorical)	plot	4 levels		
P_YOE (categorical)	plot	11 levels		
P_sp (categorical)	plot	1 level (PSMEN)		
P_thinned (categorical)	T: 295	UT: 22		
card_4wayN (categorical)	N: 49	E: 70	S:121	W: 77
card_4wayNE (categorical)	NE: 60	NW: 56	SE: 105	SW: 96
card_8way (categorical)	n: 19	ne: 32	e: 40	se: 53
	s: 53	sw: 49	w: 38	nw: 33

2.5.2 Visualising explanatory variables alone

Visualisation of explanatory variables was undertaken in R (R Core Team, 2019), using the *base* (R Core Team, 2019), *ggplot2* (Wickham, 2016) and *corrplot* (Wei & Simco, 2017) packages. First, all variables at the individual tree level and at the plot level were visualised with a simple index graph to check for anomalous values that might indicate data errors (these graphs have not been presented).

Second, variables that describe individual trees (tree height, tree diameter at breast height, tree basal area, tree live/dead status, tree top status, tree slenderness, occurrence of first forks, heights of first forks, occurrence of second forks, heights of second forks, tree pruning status, and tree pruned height) were visualised as histograms: these are available in Appendix 6.5.1.

Third, variables that describe plots were visualised as histograms (continuous and count variables) or bar charts (categorical variables). These include two plot description variables (*P_YOE* and *P_YOM*), all tree description variables, all silvicultural history variables, all topographic variables, and all weather variables. These figures are available in Appendix 6.5.2.

Fourth, numeric variables that describe plots were displayed as correlation plots, with one plot per species, to show the strength of relationships among the explanatory variables. These figures are available in Appendix 6.5.3.

2.5.3 Visualising response and explanatory variables together

Next, the three response variables - mean height of broken trees per plot (*P_BRKN_ht_mean*), proportion of damaged (broken plus horizontal) trees per plot (*Tops_prpn_DAM*), and proportion of live trees per plot (*Prpn_LIVE*) – were compared with the potential explanatory variables. First, correlation plots were calculated for the response and explanatory variables on a species-by-species basis, to explore relationships among the variables. These plots are available in Appendix 6.5.4.

Third, classification and regression trees (CART) relating the three response variables to the explanatory variables were created, to explore potentially important relationships between response and explanatory variables, and also interactions among them. Because classification and regression trees created in the *tree* package drop incomplete observations, some variables with many NA entries were omitted; otherwise, too many observations were lost. These included *P_Fk_1_ht*, *P_Fk_2_prpn*, *P_Fk_2_ht*, *Age_thin*, *Age_LastP*, *T_P_gap*, *Sph_drop*, and *P_tree_ht_mean_BRKN* (except when it was the response variable). The classification and regression trees are given in Appendix 6.5.5.

2.6 Statistical model creation

2.6.1 Establishing a difference between the species

Field observations of breakage at Geraldine Forest have suggested that Douglas-fir suffers less severe top breakage and related damage than radiata pine; therefore, statistical tests were applied to establish whether there is a significant difference between species in the per-plot mean values of the response variables, where a p-value of less than 0.05 was considered to indicate a significant difference. The difference between the mean height of broken trees per plot (*P_BRKN_ht_mean*) for each species was tested with a t-test (*P_BRKN_ht_mean* is normally distributed). The difference between the proportion of damaged trees per plot (*Tops_prpn_DAM*) and proportion of live trees per plot (*Prpn_LIVE*) for each species, which are not normally distributed, were tested with Wilcox Rank sum tests.

2.6.2 Allowing for hierarchy in the data

There is a definite hierarchy in this study's data. Data may have been generated at the individual tree level (bottom of the hierarchy); the plot level (middle of the hierarchy); or at the stand level (top of the hierarchy). All variables and all analyses in this study are at the plot level, which is useful and familiar to a forestry audience. However, in recognition of the hierarchy, some variables representing stand-level data (*P_stand*, *P_YOM*, *P_YOE*) have been trialled as random effects (i.e. grouping variables) in mixed-effects regression models, to see whether that improved predictive power over non-mixed regression models, by capturing the effects of groups that do not have explicit representation in the explanatory variables.

Grouping variables created for use in mixed-effects models should represent some similarity among the group members. The variable *P_YOE*, plot year of establishment, might capture similarities of tree genetics and treestock quality, which are likely to be the same (by species) for any given planting year; might capture the effects of tree establishment on subsequent growth, as establishment practices and weather at planting time may vary between years; and might give greater detail about the effects of weather events, which are unevenly spread in time, so that a tree planted in year 'A' might experience more unfavourable weather events than an otherwise similar tree of the same age that was planted in year 'B'. *P_stand* might have similar properties to *P_YOE,* and in addition give greater detail about the effects of planted stocking and silviculture, which are (by definition) the same across a stand. *P_YOM*, plot year of measurement, might capture the influence of differences in tree measurement technique, as measurement personnel and their skill level may vary between years. *P_YOM* is not available for Douglas-fir models: it has four levels, but a minimum of five are required to support the mixed-effects calculations.

The variable *Plot_no* offers a way of compensating for over-dispersion in logistic regression models, by use as an observation-level random effect. Note for the random forest models implemented in this study, one can include *P_YOE*, *P_stand*, and *P_YOM* as classification variables, which is analogous to the use of *P_YOE*, *P_stand*, and *P_YOM* as mixed effects, but there is no analogy to the use of *Plot_no* as an observation-level random effect.

2.6.3 Development of the fitting and validation datasets

Before creating predictive models, the data were split into fitting and validation datasets. Models were trained on the fitting dataset, and the validation set was retained to assess how well models performed when presented with previously unseen data. The R base package function *sample* was used to create a random choice of 169 from 942 (17%) of the plot identifiers, which were then excluded from model-building, and retained for model validation. The identities of the fitting and validation plots are given in Appendix 6.3. These dataset split was used for all models in this study.

2.6.4 Modelling strategies

2.6.4.1 Assumptions when creating the response variable Tops_prpn_DAM

Of the original 942 plots providing data for this study, only the plots from LiDAR survey ground control (190) and the permanent sample plots (9) included an assessment of the top status of every tree. The balance of 743 inventory plots record only a sub-sample of tree top statuses. This presents a problem for the calculation of the response variable proportion of damaged tops per plot (*Tops_prpn_DAM*), as there were three possible interpretations of the data. Either all the broken and horizontal trees in a plot were recorded, and the balance of trees all have normal tops; or only some of each were recorded, and the ratio of recorded and horizontal tops to recorded normal tops is a fair indication of the proportion across the whole plot; or only some of each were recorded, and the ratio of recorded damaged and horizontal tops to recorded normal tops is a biased indication of the proportion across the whole plot.

The forest inventory practice specified by Port Blakely, the owner of Geraldine Forest, has variability, with different minimum numbers of normal-top trees specified to be recorded and measured for height, depending on the total number of heights required for that particular survey. The inventory procedure does not specify to record the presence of every broken or horizontal tree; if it did, then all non-broken, non-horizontal tree tops could be assumed to be normal. Rather, the procedure should be interpreted as giving directions to choose a certain number of height trees, and if the trees chosen are unsuitable due to being broken or horizontal, record this and move on to a tree that has a normal top. Unfortunately, this could lead to a biased proportion, depending on how exactly the inventory crew selects the replacement height trees.

A check of the data for inventory plots shows that the minimum proportion of assessed tops (including broken, horizontal, and normal) in a plot is zero, the median is 0.28, and the mean is 0.31. Clearly, there is the opportunity for biases to arise when the proportion of tree tops assessed is relatively low in this manner. The plots arising from LiDAR and PSP, where all tops were assessed, gives the opportunity to check whether the proportion of damaged to undamaged trees in inventory plots, as estimated from known tops only, is a fair reflection of the overall plot proportion. There being no reason to suspect a different rate of damage between the two types of plot, then if the overall proportion of damage in inventory plots is similar to overall proportion of damage in LiDAR and PSP plots, then the data from inventory plots are likely to be useable. The figures for these scenarios are shown in Table 2-12, below.

Table 2-12: proportion damaged trees (mean of all plot proportions) under different assumptions.

	scenario	radiata pine	Douglas-fir
LiDAR plots and PSPs	scenario one: no assumption – all tops assessed	0.38	no data
inventory plots	scenario two: assuming all damaged tops in each plot were recorded	0.18	0.07
	scenario three: assuming proportion damaged/not from sub-sample of recorded tops is fair reflection of reality	0.45	0.22

Clearly, the proportion of damaged tops under scenario three is most like the proportion of damaged tops under scenario one. Therefore, damage proportions arising from the sub-samples by plot of tree top status have been used as the replacement for the true damage proportions, which are unknown in the case of inventory plots.

This technique has some error, however. A comparison of frequency distributions for damaged proportions from scenario one (Figure 2-2) and scenario three (Figure 2-3) for radiata pine reveals clustering for scenario 3. Clustering around zero may legitimately represent plots with zero damage. However, clustering at other points, such as 0.25, 0.33 and 0.5, is caused by the presence of proportions calculated from small-value whole numbers, for example one top broken from four measured.

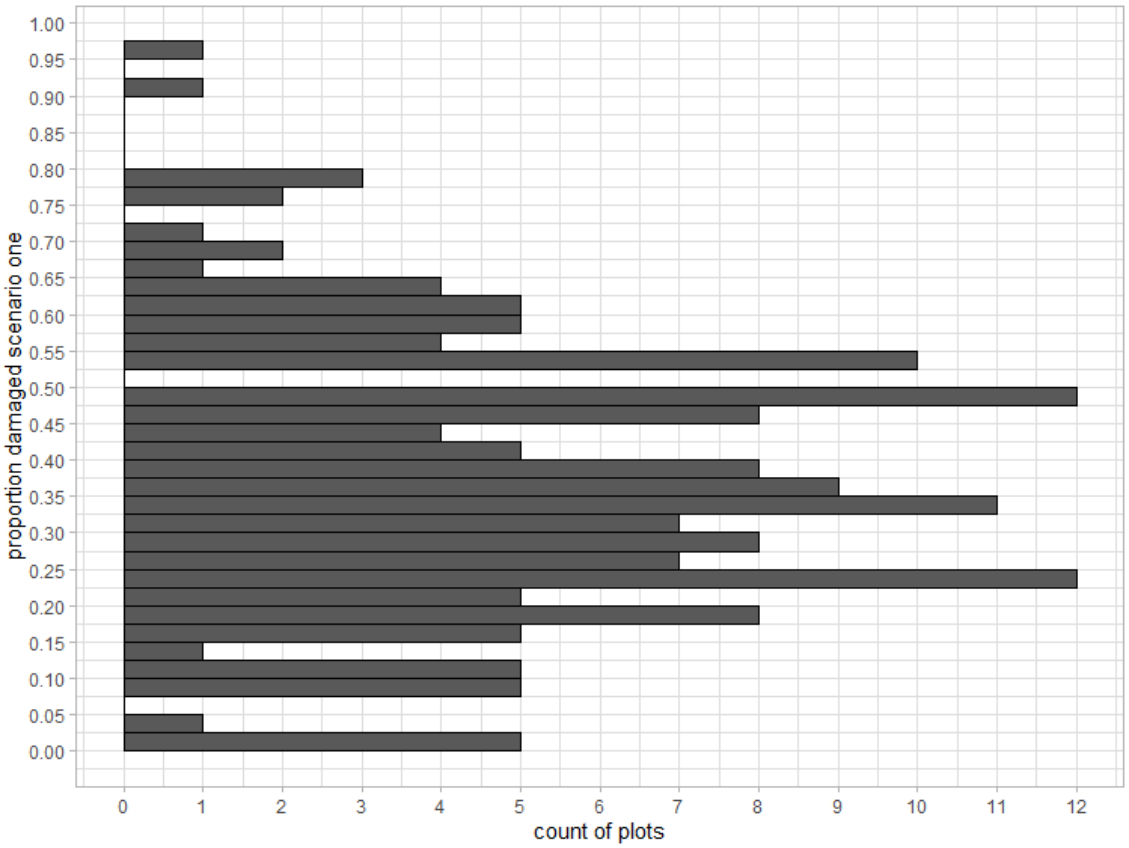

Figure 2-2: frequency distribution for proportion of trees damaged per plot, as for scenario three in Table 2-12.

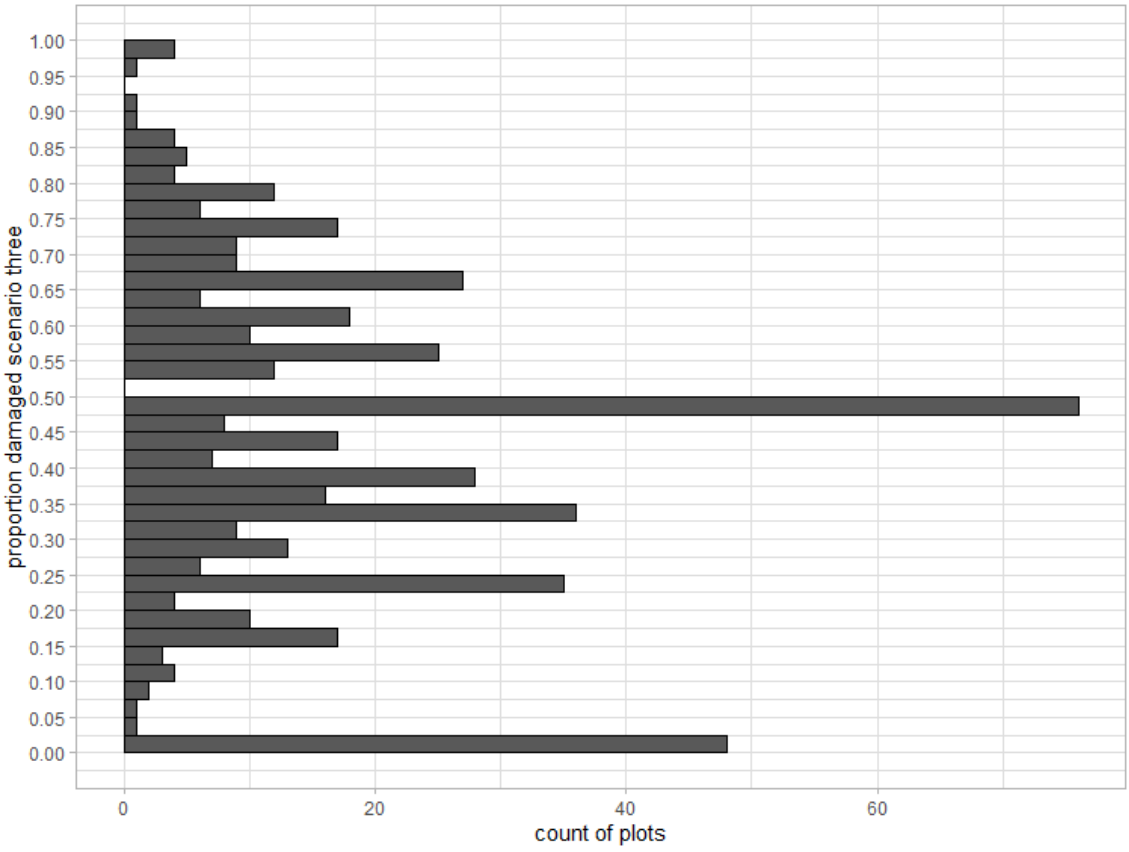

Figure 2-3: frequency distribution for proportion of trees damaged per plot, as for scenario one in Table 2-12.

This choice to use sub-samples to estimate the proportion of damaged trees is the single biggest assumption in this study. Without this assumption, however, the data available for modelling the broken proportion for radiata pine would have been severely restricted, and non-existent for Douglas-fir. To illustrate the implications of the assumption as far as possible, models of damaged proportion for radiata pine have been created separately from LiDAR and permanent sample plot data, and from all data together.

2.6.4.2 Variables discarded

It quickly became apparent that *P_Fk_1_ht* and *P_Fk_2_ht* would have to be discarded as explantory variables. Both model types used in this study, regressions and random forests, only utilise input data that is complete for each plot. For on *P_Fk_1_ht* and *P_Fk_2_ht*, many missing values arise because plots commonly have no forked trees. The true value for the average forking height for a plot with no forks must be NA – but the presence of the NA removes the entire plot from the modelling set. While *P_Fk_2_prpn* has a numeric true value (0) in the absence of second forks, only 55 radiata pine plots and one Douglas-fir plot had any second forks. The predictive contribution for these types of model of a variable that is mostly zero is limited, and so *P_Fk_2_prpn* was dropped.

2.6.4.3 Regression analysis

In this study, the nature of the response variables suggested suitable types of regression models. Table 2-13, below, shows the regression types used.

Table 2-13: types of regression analysis considered, by response variable.

response variable	distribution of response variable	range of response variable	simplest regression attempted	more complex regressions attempted
P_tree_ht_mean_BRKN	normal	radiata pine: 2.6 m – 28.1 m Douglas-fir: 3.8 – 28.3	multivariate linear regression	multivariate linear regression with mixed-effects
Tops_prpn_DAM	not normal: proportion	radiata pine: 0 - 1 Douglas-fir: 0 – 0.833	multivariate logistic regression	multivariate logistic regression with mixed-effects manual hurdle model (radiata pine only)
Prpn_LIVE	not normal: proportion	radiata pine: 0.063 - 1 Douglas-fir: 0.065 - 1	multivariate logistic regression	multivariate logistic regression with mixed-effects

The initial modelling set comprised 59 variables relating to radiata pine and 54 for Douglas-fir, where Douglas-fir has the same set as for radiata pine, less the variables related to pruning, which is not undertaken on Douglas-fir at Geraldine Forest. This leads to a large number of possible models. Therefore, regression models were built in an exploratory fashion for each species individually, following these basic steps:

1. Remove any variables confounded with the response variable, and remove from consideration variables that badly reduce the size of the data set due to missing values. *Age_thin*, *Sph_drop*, *T_P_gap* and *Final_sph* have many missing values, due to poor silviculture records for Geraldine forest.
2. Graph all (remaining) explanatory variables against the response variable, as scatterplots (for numeric and count variables) or as bar plots (for categorical variables).

3. Also, create a correlation matrix for the response variable and explanatory variables.
4. Create a model with all explanatory variables that had a correlation of >0.1 with the response variable. Assess the significance (p-values less than 0.1 were considered promising at this screening step), the collinearity (variance inflation factors of less than four were considered promising at this screening step). This model will probably be overfitted: it is a first-pass screening step.
5. Create a model with low multicollinearity and a good apparent balance between explanatory power and bias, on a trial-and error basis, attempting to capture variables that are significant in the first-pass model, and correlations apparent in the correlation matrix and graphs. Tree description variables, silvicultural history variables, and topographic variables are trialled as fixed effects, the stand-level categorical variables *P_YOM, P_YOE*, and *P_stand* are trialled as mixed effects by random intercepts.
6. Consider the explanatory power of the model fit statistics. Consider the plausibility of the model, especially the model coefficients, by domain knowledge of forestry. Consider the degree of autocorrelation in the model residuals. Consider the fit to test data.
7. For logistic regression, test model as at 6 for over-dispersion, and if present, attempt to compensate for it by adding *Plot_no* as an observation-level random effect.
8. Repeat steps 6 and 7 until a satisfactory model is created, or it is concluded that a satisfactory model cannot be created.

All numeric potential explanatory variables were centred around their means, so that the mean was zero, to reduce the collinearity between any interactions and their component variables, and to make parameter estimates more interpretable. This was particularly so for interpretation of model intercepts, because many of the candidate explanatory variables have no zero in their data range, and therefore an intercept when all explantory variables are zero, as would be calculated from uncentred variables, is nonsensical. Numeric potential explanatory were also scaled by their standard deviation, to make the units all one standard deviation. Centring and scaling are particularly recommended for mixed-effects models (Harrison et al., 2018), which have been used in this study.

Over-dispersion is a known issue in binomial logistic regression (Harrison, 2015). It occurs when the response variable has a variance in excess of the variance expected for model type, and in binomial models of proportion over-dispersion reveals itself when models have residual deviance higher than the residual degrees of freedom. The over-dispersion scale factor is calculated as residual deviance/residual degrees of freedom, except for models with mixed-effects from the stand level, where it calculated by the method described in Bolker (2019). If over-dispersion was detected, defined as scale factors >1.25, it was compensated for by using the plot number, a unique identifier, as an observation level random effect in a mixed-effects model, as described in Harrison et al. (2018).

As with all linear regressions involving categorical variables, the model outputs of regressions created in this study do not show the first level of the category, because it is by definition included in the intercept of the model, and other levels of the category then modify that intercept. The default in R, followed in this study, is to use the category with the name nearest the start of the alphabet as the base level. Table 2-14, below, shows what level of a categorical variable is the base level.

Table 2-14: base levels and listed levels in models including categorical variables.

variable	base level – does not appear in model outputs	other levels
card_4way_N	east (E)	north (N), south (S), west (W)
card_4way_NE	north-east (NE)	south-east (SE), south-west (SW), north-west (NW)
card_8way	east (e)	north (n), north-east (ne), south-east (se), south (s), south-west (sw), west (w) or north-west (nw)
P_pruned	pruned (P)	unpruned (UP)
P_thinned	thinned (T)	unthinned (UT)

2.6.4.4 Random Forests

In this study, random forest models provide a comparison with regression analysis. Random forest models were created using the *cForest* function from the *party* package in R (Hothorn, Bühlmann, Dudoit, Molinaro, & Van Der Laan, 2006; Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008; Strobl, Boulesteix, Zeileis, & Hothorn, 2007). *cForest* is an implementation of the random forest algorithm that uses conditional inference trees as its base learners. *cForest* was chosen because it is designed to mitigate bias due to factor variables with differing numbers of levels, and due to numeric data scales of different orders of magnitude, both of which are present in this study's dataset. Data were split into fitting and validation data sets, as described in section 2.6.3, and a random seed of 231 was used across all random forest models. A thousand trees were predicted for all models. The number of variables to try at each split (*mtry*) was set to the nearest rounded-up integer value of the square root of the number of variables.

By the nature of random forests, the human researcher does not choose which variables to include in a random forest model: rather, the random forest algorithm chooses from among a set of variables provided. However, the researcher can restrict the members of the set of variables provided to the random forest algorithm. There are three types of variable set used to calculate random forests in this study. This is a deliberate technique to explore whether random forests models would choose the same explanatory variables as were chosen (by the human researcher) for inclusion in the regression models, and if not, how different the sets would be. The three sets are 1) all explanatory variables; 2) the best ten explanatory variables from results of the all-variables model, 3) the same explanatory variables as for the corresponding regression.

Centring and scaling of variables is not required for random forests; but in all other respects, the variables are the same as those used for regression analysis. Creation of the random forest models in this study followed these steps:

1. Remove any variables confounded with the response variable, and remove from consideration variables that badly reduce the size of the data set due to missing values. These are the same as for the regression analyses, namely *Age_thin*, *Sph_drop*, *T_P_gap* and *Final_sph*, and the same comments apply as for the regression analyses.
2. Split data into fitting and validation sets.
3. Create an initial model (Model 1) with all potential explanatory variables from the fitting set. Assess the importance of variables, by creating a ranked table of increase in mean square error, and assess the model's fit to test data.
4. Create a refined model (Model 2) that includes the ten most important variables from results of the all-predictors model.

5. Create a refined model (Model 3) that includes the same variables that were important in the corresponding regression analysis.
6. For the best model from steps 3 -5, use that model and the test data to predict the response variable. The best model has the best apparent balance between predictive power and bias. Best models are presented in sections 3.3.2, 3.4.2, and 3.5.2: fit statistics for the other two random forest models of any given species/response variable combination may be found in Appendix 6.7.

Random forests generate slightly different results for each model run. To maintain consistency, results (whether numeric or graphical) reported for each random forest model were taken from a single run of the model.

2.6.4.5 Dealing with correlation among predictors

Variables describing the dimensions of trees are often highly correlated with each other. For example, for a given sample of trees, the diameter at breast height of trees is likely to be strongly correlated with the basal area, which is partly calculated from diameter. Similarly, there is likely to be a correlation between tree age, and tree height, as older trees have grown for a longer period and will probably have become taller. This introduces multicollinearity among potential explanatory variables. The degree of multicollinearity is displayed in Appendix 6.5.3; note that this is multicollinearity for the entire data set, not the model fitting data alone.

Multicollinearity among potential explanatory variables is a problem for regression analysis, because correlated explantory variables address some of the same variance in the response variable, meaning the effects of each cannot be independently established. It is wise to check for multicollinearity, and unwise to include in a model several variables addressing broadly the same idea (Harrison et al., 2018). Therefore, models created during regression analysis include at most one variable from each of these groups:

- measurement age, weather variables (which depend on measurement age) and mean height of trees with normal tops
- the variants of morphometric protection index
- the variants of aspect
- any measure of aspect and either of *WindShelt_NE1* or *WindShelt_S1*
- any measure of MPI and either of *WindShelt_NE1* or *WindShelt_S1*
- basal area and the combination normal-top tree diameter plus stocking

To assess the degree of multicollinearity, all regression models had their variance inflation factors (VIFs) calculated, using the *vif* function in the R package *lme4*. The *vif* function performs, for each variable included in a linear or generalised linear model, a regression of that variable by all of the other variables, and returns $1/(1-R^2)$. For any given VIF, subtract one, multiply the figure remaining by 100, and the result is the percentage inflation of the variance of the coefficient. A VIF of 1 means no collinearity. A value of 1.6 means that the variance of a model coefficient is 60% larger than it would have been in the absence of multicollinearity. Variable combinations with VIFs of greater than 2 were not used during modelling.

Random forests are a non-parametric modelling strategy, which in principle can handle a high number of predictors relative to observations, complex interactions, and highly correlated predictor variables (Strobl et al., 2008). Random forest models were used partly because these characteristic are all present in the data for this research.

2.6.4.6 Checks for autocorrelation in model results

As a check on the desirability (or otherwise) of models, the R package 'ape' (Paradis & Schliep, 2018) was used to calculate the Moran statistic to test for autocorrelation in the models' residuals. A model that adequately explains spatial autocorrelation (similarity imposed by geographical proximity) between data points by including spatial variables will not have autocorrelated residuals. Likewise, a model that adequately explains first-order conditional autoregressive structure (similarity imposed by membership of a group) by including grouping variables will not have autocorrelated residuals. The other type of autocorrelation, temporal autocorrelation, is not an issue for this research, as each plot has only one measurement date in the data set. As the co-ordinates of all study plots are known on a Cartesian grid (the New Zealand Transverse Mercator map projection), the weights matrix for the Moran statistic was defined as the inverse of a matrix of the Euclidean distance between each point.

2.6.5 Interpretation of the model outputs

2.6.5.1 Model fit statistics

The statistics given for the fit of models to fitting data are: R^2 (further explained below); mean absolute percentage error (MAPE), chosen instead of root mean square error because the proportion of damaged trees per plot (*Tops_prpn_DAM*) and proportion of live trees per plot (*Prpn_LIVE*) inherently range from 0 – 1; and the slope and intercept of a bias check, which is an ordinary least-squares regression of the predicted values for the fitting data by the actual values of the fitting data. In addition, the result of the Shapiro-Wilk test for normality of residuals is given for linear regressions, and the over-dispersion statistic is given for logistic regressions.

Many variants of R^2 exist in the statistical literature, so some discussion of the R^2 used when fitting models to the fitting data must be discussed. The R^2 values used for non-mixed models are adjusted R^2 from the model summary of package *lme4* (Bates, Mächler, Bolker, & Walker, 2015) for multivariate linear regressions, and McFadden's pseudo-R^2 from the function rsquared from the package *piecewiseSEM* (Lefcheck, 2016) for multivariate logistic regressions. Although Nakagawa, Schielzeth, and O'Hara (2013) created a means of estimating marginal (fixed-effects) and conditional (full-model) R^2 for mixed-effects models, Harrison et al. (2018) note that the conditional R^2 is not useful for models containing an OLRE. The mixed-effects logistic regressions in this research do contain *Plot_no* as an OLRE. The alternative metric chosen for mixed-effects logistic regressions is to present the simple (unadjusted) R^2, which is the square of the correlation coefficient, of the relationship between real and predicted values of the fitting data, and to present alongside it the McFadden's pseudo-R^2 of a model that is equivalent in formulation, *except that* it omits the mixed effects. The simple R^2 will probably be too high, as it has not been adjusted downwards by the number of variables in the model, and the McFadden's pseudo-R^2 will probably be too low, because the R^2 will not be raised by the 'shrinkage' phenomenon of mixed-effects models. However, the comparison should give an indication of the relative contribution of adding the OLRE to the models. For random forest models, the R^2 from the function cforestStats in the package *caret* was used: this also returns the simple R^2.

The model statistics presented for model validation are of the same types: R^2, MAPE, slope and intercept of the bias check. For test data, R^2 is again the simple R^2, the squared value of the correlation between actual values for test data and predicted values for test data. For both train and test data, statistics for a perfect fit would be $R^2 = 1$, MAPE = 0, bias check slope = 1, bias check intercept = 0, over-dispersion factor = 1.

2.6.5.2 Interpretation of logistic regression model coefficients

Logistic regression is the correct technique when the response variable is formulated as 'm successes, n failures', also known as a binomial count. When considered at the plot level, *Tops_prpn_DAM* (the proportion of damaged trees per plot) and *Prpn_LIVE* (the proportion of live trees per plot) are binomial counts, comprising a count of damaged trees ('successes'), and a count of undamaged trees ('failures'), or live trees ('successes') and dead trees ('failures). However, logistic regression outputs are not intuitive to interpret in their raw form. If p = the probability of some event occurring, and because for binomial counts the *probability* of success is the same as the *predicted proportion* of successes, then instead of the model form familiar from linear regressions:

$$y = a + b_1x_1 + \ldots + b_nx_n$$

the model form for logistic regression is

$$log(p/1-p) = a + b_1x_1 + \ldots + b_nx_n$$

In words, this says the response to the linear predictors $x_1 \ldots x_n$ is the natural log of the *odds* of success occurring, given those predictors. The odds are the $p/1-p$, or the probability of success divided by the probability of failure. As logarithms are difficult to interpret directly, we can apply, for the predictor variable coefficients $b_1 \ldots b_n$

$$(exp(b)-1)*100$$

which gives the percentage change in odds for the response variable, given a change of one unit (here one standard deviation) in the predictor variable, with the direction of the change indicated by the sign. This figure has been given alongside the raw model coefficients for logistic regression in the Results, except for model intercepts, where it is not a meaningful figure.

2.6.6 Modelling the plot mean broken height

2.6.6.1 Plot mean broken height using multivariate linear regression

The response variable *P_BRKN_ht_mean*, the mean height of tree breakage for each plot, is a continuous numeric variable. Density plots of *P_BRKN_ht_mean* for both radiata pine and Douglas-fir showed no skew either way, and the Shapiro-Wilk test showed the response variable *P_BRKN_ht_mean* was not significantly different from a normal distribution ($p = 0.3375$ for radiata pine and $p = 0.3518$ for Douglas-fir).

Therefore, *P_BRKN_ht_mean* was first modelled with multivariate linear regression, using the linear model (*lm*) function from the R package *lmtest* (Zeileis & Hothorn, 2002) and the modelling process outlined in section 2.6.4.3. Model residuals were inspected graphically at each step of model development. The use of Mallow's cP as a model selection tool was considered, but Mallow's cP does not take account of collinearity among its inputs. Therefore, graphs of the Mallow's cP statistic were used to suggest, but not decide, well-performing model combinations.

Next, the best linear regression was modified to a mixed-effects framework, where the stand-level variables *P_stand*, *P_YOE* and *P_YOM* were trialled as intercept-only mixed effects, using the *lmer* function from the *lme4* package in R (Bates et al., 2015). A preferred model was chosen for each species, based on model fit statistics.

2.6.6.2 Plot mean broken height using Random Forests

Random forest models of the per-plot mean broken height damaged were created for each of radiata pine and Douglas-fir, using the *cForest* function from the *caret* package in R (Hothorn et al., 2006; Strobl et al., 2008; Strobl et al., 2007). Three random forest models of each species were created and tested, following the three scenarios for variable inclusion given in 2.6.4.4. The preferred model was then applied to validation dataset the using *predict* function of the R base package.

2.6.7 Modelling the proportion of damaged trees

2.6.7.1 Plot proportion of damaged trees per plot using logistic regression

The response variable *Tops_DAM_prpn*, the proportion of trees with tops damaged (trees with broken tops plus horizontal trees), is a proportion variable and so is appropriate for modelling with binomial logistic regression, using the *glm* (generalised linear model) function available in the R package *lme4* (Bates et al., 2015), with a logit link. As the response is a proportion, not a binary response, a two-vector response variable was used, containing the counts of damaged and undamaged trees per plot. There are two variant for radiata pine proportion damaged per plot: a model including all plots, and a model including only the plots where the top status of every tree is known.

As a variant of the above, the best logistic regressions were modified to explore the usefulness of including the stand-level variables *P_stand* (stand number), *P_YOE* (plot year of establishment) and *P_YOM* (plot year of measurement), which were identified (section 2.6.2) as being hierarchical to the plot-level variables, and to explore the usefulness of including the observation-level random effect (OLRE) *Plot_no*. These variables were included as intercept-only random effects in a mixed-effects modelling framework, using package *lme4* (Bates et al., 2015).

2.6.7.2 Plot proportion of damaged trees per plot with hurdle models

A hurdle model is a two-part model, which include a probabilistic model of whether some event occurs or not, and a regression model of the severity of the event when it occurs. This fits conceptually for the response variable *Tops_prpn_DAM* (the proportion of damaged trees per plot), where the question may be posed: *if 1): only some of the plots have damage, then 2): what is the damaged proportion in plots that do have damage?*

Two attempts were made to use hurdle models, both with radiata pine data for all plots. The first attempt was to use the zero-adjusted binomial (ZAB) model from the R package *VGAM* (Yee, 2019), which is a hurdle model. The ZAB model uses logistic regression on a binary outcome variable for the first part, and logistic regression on a proportion variable for the second part. All the attempted formulations of the ZAB model had a common feature: the regression model was intercept-only, indicating the available explantory variables could not address 2*): what is the damaged proportion in plots that do have damage?* Consequently, no results have been reported for the attempted uses of the ZAB model.

Second, a manual hurdle model was created, comprising

- a binary logistic regression on all data for 1) as above,
- a separate multivariate logistic regression only for plots with some damaged for 2) as above
- multiplication of the 0/1 response of 1) by the proportion response of 2)

This hurdle model was manual in the sense that separate models were created for each of the probabilistic and regression models, and then the outcomes were multiplied together. This is a similar approach to that taken by Fletcher, MacKenzie, and Villouta (2005), in their research into modelling ecological abundance data skewed by large numbers of zeros.

2.6.7.3 Proportion of damaged trees per plot using Random Forests

Random forest models of the per-plot proportion damaged (*Tops_prpn_DAM*) were created using the *cForest* function and tested with the *predict* function from the *cForest* package in R (Hothorn et al., 2006; Strobl et al., 2008; Strobl et al., 2007). Three random forest models were created and tested, following the three scenarios for variable inclusion given in 2.6.4.4. In the same manner as for logistic regression, there are two random forest models of *Tops_prpn_DAM* – models including all plots, and models including only the plots where the top status of every tree is known.

2.6.8 Modelling the proportion of live trees

2.6.8.1 Plot proportion of live trees per plot using logistic regression

The response variable *Prpn_LIVE*, the proportion of live trees in a plot, is a proportion variable and so is appropriate for modelling with binomial logistic regression, using the generalised linear model (*glm*) function from the R package *lme4* (Bates et al., 2015), with a logit link. As the response is a proportion, not a binary 0/1 response, a two-vector response variable was used, containing the counts of live and dead trees per plot.

In the same manner as for *Tops_prpn_DAM*, the best initial binomial regression was modified to explore the usefulness of including the stand-level variables *P_stand*, *P_YOE* and *P_YOM*, which are hierarchical to the plot-level variables, and also the ORLE *Plot_no*. Their inclusion was as intercept-only random effects in a mixed-effects modelling framework, using package *lme4* (Bates et al., 2015).

2.6.8.2 Plot proportion of live trees per plot using Random Forests

Random forest models of the proportion of live trees (*Prpn_LIVE*) were created using the *cForest* function and tested with the *predict* function from the *cForest* package in R (Hothorn et al., 2006; Strobl et al., 2008; Strobl et al., 2007). Three random forest models were created and tested, following the three scenarios for variable inclusion given in 2.6.4.4.

3 Results

This section details the various models created to explain relationships between the response variables plot mean height of broken trees (*P_BRKN_ht_mean*), proportion of damaged trees per plot (*Tops_prpn_DAM*) and proportion of live trees per plot *(Prpn_LIVE)*, and the predictor variables, for each species. Throughout Results, MAPE means mean absolute percentage error.

3.1 Establishing a difference between the species

Observations by field staff at Geraldine Forest suggest that Douglas-fir is less likely to exhibit damage from wind and snow than radiata pine. Therefore, tests were undertaken to establish whether there is a difference between radiata pine and Douglas-fir in the mean values of the response variables *P_BRKN_ht_mean*, *Tops_prpn_DAM*, and *Prpn_LIVE*.

The results, given in Table 3-1 below, clearly show that there are differences between the species, with Douglas-fir exhibiting higher broken heights, lower proportion damaged and a lower proportion of live trees. With those differences established, the predictive models presented in the sections that follow have been created on a species-by-species basis.

Table 3-1: tests for differences between response variables. Means or proportions, and standard errors.

test	variable	hypothesis	value for radiata pine (625 plots)	value for Douglas-fir (317 plots)	p-value	Inference
t-test, two-tailed	*P_BRKN_ht_mean*	alternative: mean of *P_BRKN_ht_mean* is different between species null: means are the same	14.5 m (0.187 m)	15.4 m (0.246 m)	0.028	mean broken heights are lower in radiata pine
Wilcoxon rank sum test, two-tailed	*Tops_prpn_DAM*	alternative: mean of *Tops_prpn_DAM* is different between species null: means are the same	0.432 (0.009)	0.221 (0.013 m)	<0.0001	mean proportion damaged is higher in radiata pine
Wilcoxon rank sum test, two-tailed	*Prpn_LIVE*	alternative: mean of *Prpn_LIVE* is different between species null: means are the same	0.983 (mean)	0.969 (mean)	<0.0001	mean proportion of live trees is higher in radiata pine

3.2 Summary of results for the species-level models

None of the models created provide high-quality descriptions of the variables modelled. For the purposes of these results and the following discussion, the model explanatory power has been described as one of three possible categories: moderate explanatory power is indicated by fit statistics of $R^2 > 0.4$ and bias test slope >0.35; low explanatory power is indicated by R^2 0.2 – 0.4 and bias test slope > 0.2 – 0.35; and very low explanatory power is indicated by $R^2 < 0.2$ and bias slope test <0.2.

With those categories in mind, and referring to the model fit statistics given Table 3-2, below, it is found that models of *P_tree_ht_mean_BRKN* are of higher explantory power than models of *Tops_prpn_DAM*, which in turn are of higher explantory power than models of *Prpn_LIVE*. As shown by their lower-valued fit statistics, Douglas-fir models have lower explanatory power than radiata pine models, and random forest models generally have lower explanatory power than regression analyses, although the random forest for radiata pine *P_tree_ht_mean_BRKN* is an exception because it has fit statistics better than its regression equivalent.

Table 3-2, below, summarises the fit statistics for each model for the fitting and validation datasets. More complete reporting of results may be found in sections 3.3, 3.4, and 3.5, and also Appendix 6.7. Please refer to section 2.6.4 for a description of the criteria used when deciding upon variables contained in models.

Table 3-2: Comparison of model performance on fitting and validation data.

variable	model type	species	data	R^2 (full model)	R^2 (fixed effects)	MAPE	bias: intercept	bias: slope	better than random forest
P_tree_ht_ mean_BRKN	regression	radiata pine	fitting	0.426	0.372	24.5	8.919	0.386	no
			validation	0.414	*(na)*	23.7	8.714	0.407	no
		Douglas-fir	fitting	0.396	0.335	25.6	9.552	0.378	yes
			validation	0.419	*(na)*	33.0	9.890	0.383	yes
	random forest	radiata pine	fitting	0.422	*(na)*	25.2	8.902	0.392	*(na)*
			validation	0.527	*(na)*	20.2	7.949	0.477	*(na)*
		Douglas-fir	fitting	0.171	*(na)*	25.1	12.724	0.172	*(na)*
			validation	0.295	*(na)*	21.5	12.115	0.208	*(na)*
Tops_prpn_ DAM	regression	radiata pine (all plots)	fitting	0.579	0.113	11.7	0.275	0.429	yes
			validation	0.121	*(na)*	15.9	0.379	0.198	no
		radiata pine (hurdle model)	fitting	*(na)*	*(na)*	*(na)*	*(na)*	*(na)*	*(na)*
			validation	0.179	(na)	14.6	0.369	0.254	***(na)***
		radiata pine (full tops assessments)	fitting	0.799	0.382	6.8	0.139	0.641	yes
			validation	0.545	(na)	13.9	0.253	0.311	yes
		Douglas-fir	fitting	0.401	0.114	15.2	0.184	0.340	yes
			validation	0.283	*(na)*	16.4	0.186	0.321	yes
	random forest	radiata pine (all plots)	fitting	0.267	*(na)*	15.5	0.333	0.243	*(na)*
			validation	0.259	*(na)*	15.2	0.339	0.224	*(na)*
		radiata pine (full tops assessment)	fitting	0.226	*(na)*	13.0	0.292	0.262	*(na)*
			validation	0.325	*(na)*	14.7	0.292	0.261	*(na)*
		Douglas-fir	fitting	0.191	*(na)*	16.4	0.174	0.196	*(na)*
			validation	0.324	*(na)*	18.2	0.175	0.215	*(na)*
Prpn_LIVE	regression	radiata pine	fitting	0.750	0.117	1.4	0.523	0.471	yes
			validation	0.117	*(na)*	2.0	0.913	0.079	no
		Douglas-fir	fitting	0.750	0.093	1.4	0.778	0.200	yes
			validation	0.114	*(na)*	4.6	0.868	0.104	yes
	random forest	radiata pine	fitting	0.184	*(na)*	2.1	0.805	0.182	*(na)*
			validation	0.226	*(na)*	2.0	0.739	0.247	*(na)*
		Douglas-fir	fitting	0.097	*(na)*	3.3	0.864	0.108	*(na)*
			validation	0.025	*(na)*	3.8	0.925	0.041	*(na)*

(na): not applicable.

3.3 Species-level results for plot mean broken height

This section details, for each species, the models created to investigate the relationships between the response variable mean height of broken trees per plot (*P_BRKN_ht_mean)* and the explanatory variables.

When reading the results of the models of plot mean height of broken trees (*P_tree_ht_mean_BRKN)*, for both modelling methods and both species, note that the fitting and validation datasets were filtered to include only plots with one or more broken tops.

3.3.1 Plot mean broken height modelled with linear regression

Linear regression was used to model plot mean broken height (*P_tree_ht_mean_BRKN*), which is a continuous normally-distributed variable. Different types of regression were best for each species. The candidate variables and the variables included in the models are shown in Table 3-3 and Table 3-4.

As these are linear regression models, the model coefficients for continuous variables represent the change in the outcome expected for a change of one unit in that variable. Because these continuous variables are centred by the mean, and standardised by the standard deviation, one unit is one standard deviation. Categorical variable coefficients give the shift in the model intercept expected for the level of the category. Random intercepts represent the variance in the model intercept arising from membership of that random effect group.

Variables with a normal distribution, including *P_tree_ht_mean_BRKN*, do not experience over-dispersion. Therefore, the plot number (*Plot_no*), used in logistic regression (sections 3.4 and Table 3-4) as an observation-level random effect to correct over-dispersion, is not used in that manner for *P_tree_ht_mean_BRKN*, and so is not listed in the candidate variables in Table 3-3 or Table 3-4, below.

3.3.1.1 Radiata pine

The results of modelling radiata pine plot mean height of broken trees (*P_tree_ht_mean_BRKN)*, as shown in Table 3-3 below, describe a model with moderate explantory power. The model fitting R^2 is 0.426, the model validation R^2 is 0.414, the slope for the model fitting bias check 0.386, and the slope for the model validation bias check is 0.407.

Table 3-3: details of best regression model, for radiata pine P_tree_ht_mean_BRKN.

Model type	**multivariate linear regression with mixed-effects**
Candidate variables	fixed: P_age_meas_cs, P_BA_ha_equiv_cs, P_sph_equiv_cs, P_slend_mean_cs, P_Fk_1_prpn_cs, P_pruned, P_pru_prpn_cs, P_pru_ht_cs, P_slope_cs, P_alt_cs, card_4wayN, card_4wayNE, card_8way, MPI_100_cs, MPI_200_cs, MPI_500_cs, MPI_1000_cs, MPI_2000_cs, WindSheltS1_cs, WindSheltNE1_cs, P_thinned, Estab_sph_cs, u_wind_tim_cs, u_rain_cs, u_min_temp_cs, u_mint_rain_cs, u_air_pr_cs, u_rain_wind_tim_cs, P_dbh_mean_NRML_cs, P_tree_ht_mean_NRML_cs, Tops_prpn_DAM_cs, Prpn_LIVE_cs random: P_YOE, P_YOM, P_stand

Model fitting	420 observations			
Fixed effects	variable	coefficient	std. error	p-value
	Intercept	14.486	0.359	<0.0001
	P_age_meas_cs	2.747	0.258	<0.0001
	P_pru_prpn_cs	1.265	0.226	<0.0001
	P_alt_cs	-0.954	0.222	<0.0001
	card_4wayNE – NW	0.832	0.454	0.0674
	card_4wayNE – SE	-0.526	0.585	0.3685
	card_4wayNE – SW	-0.529	0.636	0.4062

Random intercepts	group	levels	variance	std. dev.
	P_stand	61	1.291	1.136

Fit statistics	marginal R^2 (fixed effects)	conditional R^2 (full model)	p-value, Shapiro-Wilk test	mean	MAPE	bias: intercept	bias: slope
	0.372	0.426	0.812 (normal residuals)	14.520	24.5	8.919	0.386

Autocorrelation of residuals	Moran's I observed	Moran's I expected	p-value	autocorrelated
	0.0041	-0.0020	0.3272	no

Model validation	94 observations				
Fit statistics	R^2	mean	MAPE	bias: intercept	bias: slope
	0.414	14.4	23.7	8.714	0.407

Plotting the actual versus predicted values for the validation set and fitting sets together (Figure 3-1) illustrates the results of using this model. The model bias is shown by the tendency of the values predicted not to follow the 1:1 line. This is also not a particularly precise model, with a wide variance in predicted values.

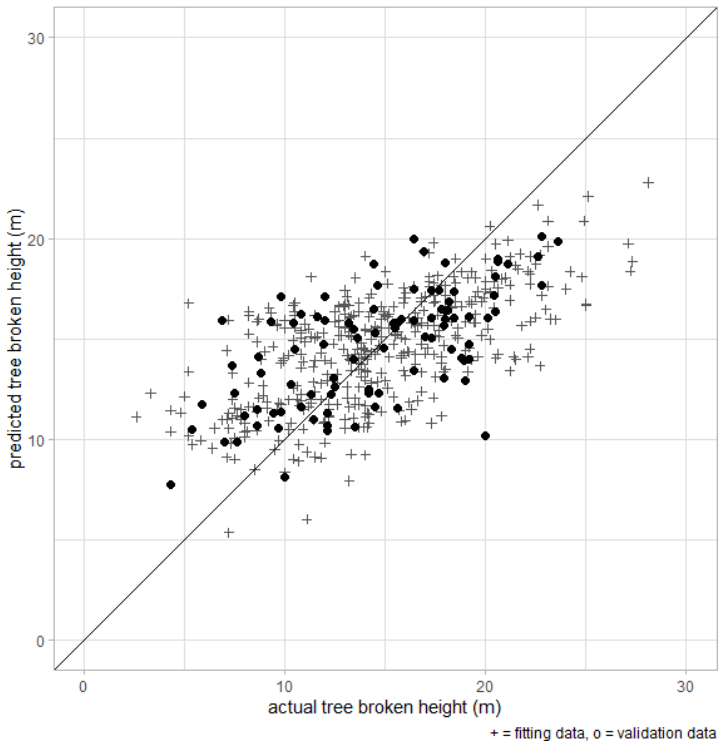

Figure 3-1: Visualising best regression model for radiata pine P_tree_ht_mean_BRKN. Actual and predicted values from validation and fitting data; 1:1 line for reference.

3.3.1.2 Douglas-fir

As may be seen below in Table 3-4, the best model for Douglas-fir plot mean height of broken trees (*P_tree_ht_mean_BRKN)* is a multivariate linear mixed-effects regression. The regression contains two categorical variables, P_thinned and card4wayN, and is dominated by P_thinned. The model fitting R^2 is 0.395, the model validation R^2 is 0.419, the slope for the model fitting bias check 0.378, and the slope for the model validation bias check is 0.383.

Although the numeric results suggest that this is a model with low R^2 (in the 0.2 – 0.4 range) and moderate bias (above 0.35), plotting the actual and predicted values (Figure 3-2) illustrates that the best model fits both fitting and validation data poorly. The domination of the regression by P_thinned reveals itself as the data sorting into two bands, where the lower band is the data for the thinned plots, and the higher band is the data for the unthinned plots.

Table 3-4: details of best regression model, for Douglas-fir P_tree_ht_mean_BRKN.

model type	**multivariate linear regression with mixed-effects**
candidate variables	fixed: P_age_meas_cs, P_BA_ha_equiv_cs, P_sph_equiv_cs, P_slend_mean_cs, P_Fk_1_prpn_cs, P_slope_cs, P_alt_cs, card_4wayN, card_4wayNE, card_8way, MPI_100_cs, MPI_200_cs, MPI_500_cs, MPI_1000_cs, MPI_2000_cs, WindSheltS1_cs, WindSheltNE1_cs, P_thinned, Estab_sph_cs, u_wind_tim_cs, u_rain_cs, u_min_temp_cs, u_mint_rain_cs, u_air_pr_cs, u_rain_wind_tim_cs, P_dbh_mean_NRML_cs, P_tree_ht_mean_NRML_cs, Tops_prpn_DAM_cs, Prpn_LIVE_cs (no prune vars) random: P_YOE, P_stand (P_YOM excluded as <5 levels)

fixed effects	variable	coefficient	std. error	p-value
	Intercept	13.982	0.530	<0.0001
	P_thinned – UT	8.152	1.633	0.0033
	P_tree_ht_mean_NRML_cs	0.619	0.347	0.0771
	Tops_prpn_DAM_cs	0.968	0.455	0.0352

random intercepts	group	levels	variance	std. dev.
	P_stand	17	0.995	0.997

fit statistics	marginal R^2 (fixed effects)	conditional R^2 (full model)	p-value, Shapiro-Wilk test	mean	MAPE	bias: intercept	bias: slope
	0.335	0.396	0.884 (normal residuals)	15.355	17.7	9.552	0.378

autocorrelation of residuals	Moran's I observed	Moran's I expected	p-value	autocorrelated
	-0.0090	-0.0185	0.779	no

model validation	35 observations				
fit statistics	R^2	mean	MAPE	bias: intercept	bias: slope
	0.419	15.218	33.0	9.890	0.383

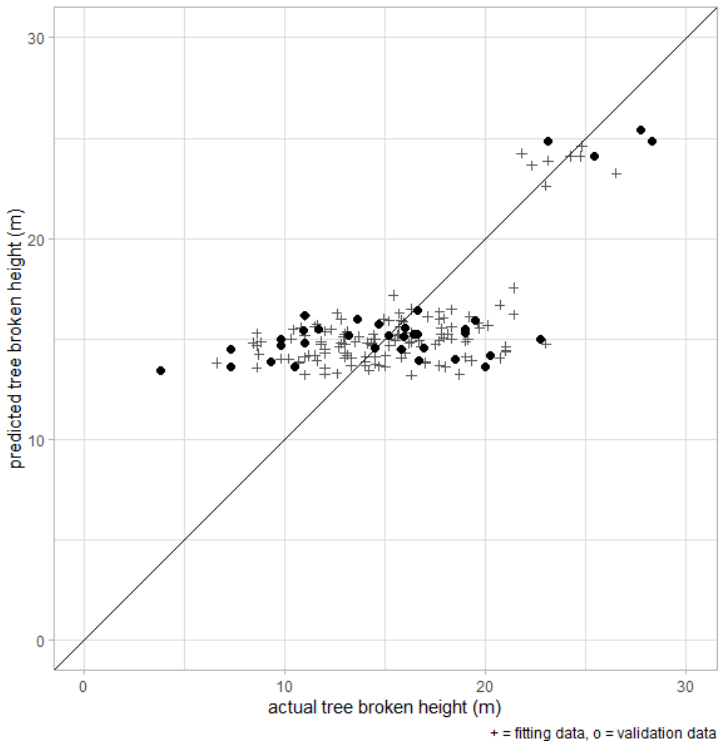

Figure 3-2: Visualising best regression model for Douglas-fir P_tree_ht_mean_BRKN. Actual and predicted values from validation and fitting data; 1:1 line for reference.

3.3.2 Plot mean broken height modelled with Random Forests

Three alternative random forest models were created for plot mean height of broken trees (*P_tree_ht_mean_BRKN),* for each of radiata pine and Douglas-fir. The three alternatives were all variables included, the top ten variables by explanatory power, and variables analogous to those used in the best corresponding regression model. Results of the best model are shown below, and the other results may be found in Appendix 6.7. As random forests do not have model coefficients, a plot of the relative importance of variables has been included.

3.3.2.1 Radiata pine

The best random forest model of *P_tree_ht_mean_BRKN* for radiata pine has moderate explanatory power. The model has these fit statistics: model fitting R^2 0.422, model validation R^2 0.527, model fitting bias check slope 0.392, and model validation bias check slope 0.477, as is shown in Table 3-5, below. These fit statistics are similar in validation and better in fitting than the corresponding regression model for radiata pine *P_tree_ht_mean_BRKN* (section 3.3.1.1).

Table 3-5: details of random forest models of radiata pine plot P_tree_ht_mean_BRKN, including model fit statistics and identification of best model.

model type	**random forest with conditional inference trees**						
candidate variables	P_age_meas, P_BA_ha_equiv, P_sph_equiv, P_slend_mean, P_Fk_1_prpn, P_pruned, P_pru_prpn, P_pru_ht, P_slope, P_alt, card_4wayN, card_4wayNE, card_8way, MPI_100, MPI_200, MPI_500, MPI_1000, MPI_2000, WindSheltS1, WindSheltNE1, P_thinned, Estab_sph, u_wind_tim, u_rain, u_min_temp, u_mint_rain, u_air_pr, u_rain_wind_tim, P_dbh_mean_NRML, P_tree_ht_mean_NRML, Tops_prpn_DAM, Prpn_LIVE, P_YOE, P_YOM, P_stand						
model fitting	420 observations						
choice of variables	variables in model		variables at each split	R^2	mean	MAPE	bias: intercept	bias: slope
top ten variables by explanatory power	P_YOM, P_age_meas, P_pru_prpn, u_wind_tim, P_tree_ht_mean_NRML, P_BA_ha_equiv, P_pru_ht, u_rain, u_air_pr, u_rain_wind_tim		4	0.414	14.525	25.3	8.856	0.396

fit statistics best model	Moran's I observed	Moran's I expected	p-value	autocorrelated
	-0.0014	-0.0026	0.8745	no

model validation	94 observations				
fit statistics best model	R^2	mean	MAPE	bias: intercept	bias: slope
	0.525	14.584	20.5	7.87	0.486

Figure 3-3, below, shows the relative importance of variables for this model.

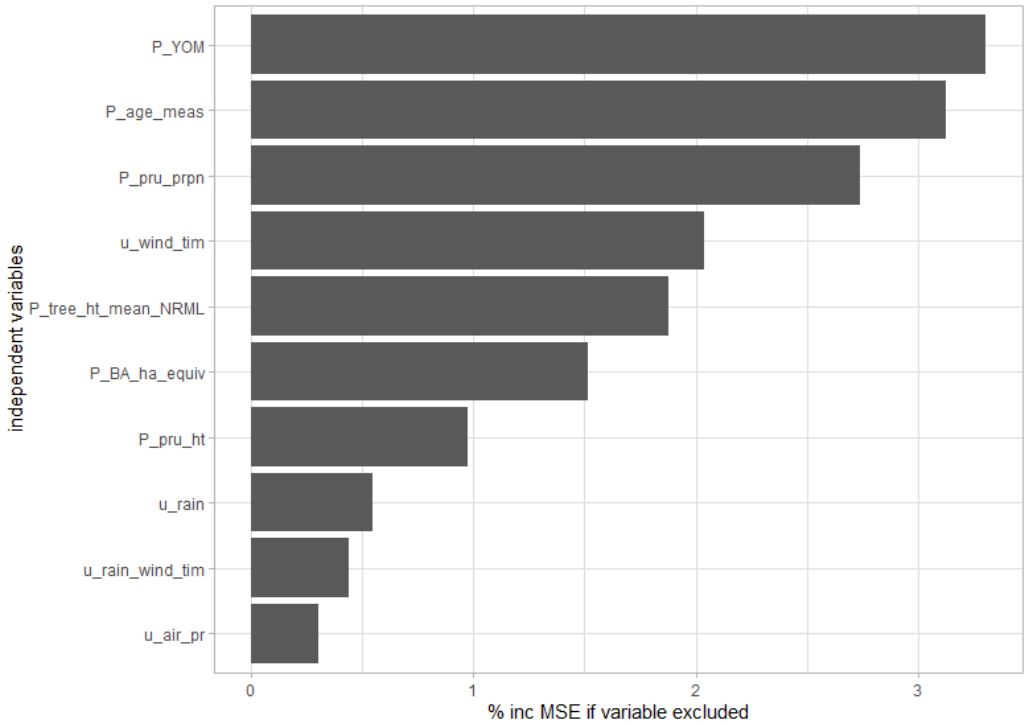

Figure 3-3: relative importance of variables for best random forest model for radiata pine P_tree_ht_mean_BRKN.

Plotting the actual versus predicted values for the validation set and fitting sets together, in Figure 3-4, illustrates that both low and high real broken heights are predicted as such, but the relationship has some bias and low precision. This is a similar patter to that shown in Figure 3-1, which plotted the results of using the corresponding regression analysis.

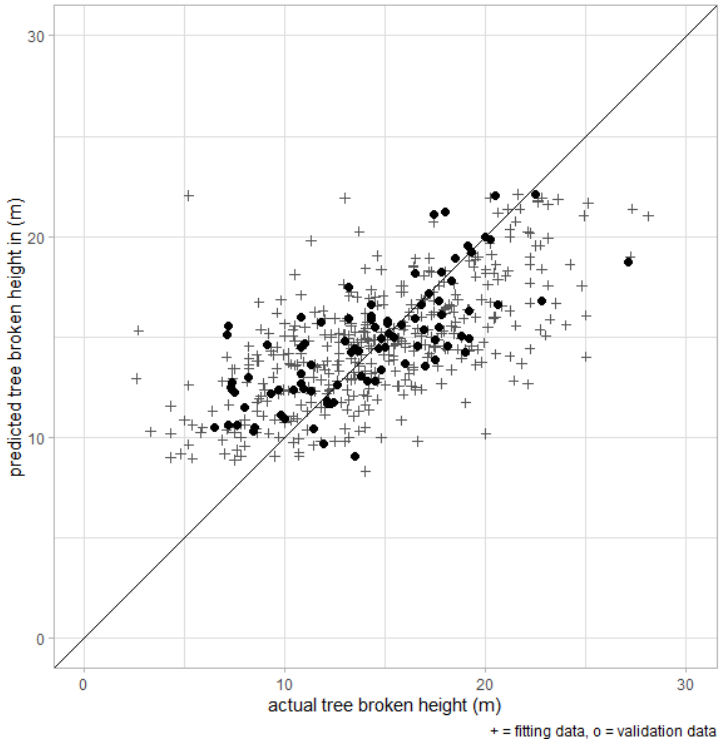

Figure 3-4: Visualising best random forest model for radiata pine P_tree_ht_mean_BRKN. Actual and predicted values from validation and fitting data; 1:1 line for reference.

3.3.2.2 Douglas-fir

The best random forest model of *P_tree_ht_mean_BRKN* for Douglas-fir has very low explanatory power. The model has these fit statistics: model fitting R^2 is 0.171, model validation R^2 0.295, model fitting bias check slope 0.172, and model validation bias check the slope 0.208, as is shown in Table 3-6, below. These fit statistics are worse in validation and fitting than the corresponding regression model for Douglas-fir *P_tree_ht_mean_BRKN* (section 3.3.1.2).

Table 3-6: details of random forest models of Douglas-fir P_tree_ht_mean_BRKN, including model fit statistics and identification of best model.

model type	**random forest with conditional inference trees**
Candidate variables	P_age_meas, P_BA_ha_equiv, P_sph_equiv, P_slend_mean, P_Fk_1_prpn, P_slope, P_alt, card_4wayN, card_4wayNE, card_8way, MPI_100, MPI_200, MPI_500, MPI_1000, MPI_2000, WindSheltS1, WindSheltNE1, P_thinned, Estab_sph, u_wind_tim, u_rain, u_min_temp, u_mint_rain, u_air_pr, u_rain_wind_tim, P_dbh_mean_NRML, P_tree_ht_mean_NRML, Tops_prpn_DAM, Prpn_LIVE, P_YOE, P_YOM, P_stand (no prune vars)

model fitting	132 observations						
choice of variables	variables in model	variables at each split	R^2	mean	MAPE	bias: intercept	bias: slope
top ten variables by explanatory power	P_thinned, P_YOE, P_tree_ht_mean_NRML, P_slend_mean, P_stand, P_YOM, P_age_meas, P_stand, u_rain_wind_tim, P_alt	4	0.171	15.351	25.1	12.724	0.172

fit statistics	Moran's I observed	Moran's I expected	p-value	autocorrelated
best model	0.0095	-0.0074	0.3426	no

model validation	35 observations				
fit statistics	R^2	mean	MAPE	bias: intercept	bias: slope
best model	0.295	15.340	21.5	12.115	0.208

Figure 3-5, below, shows the relative importance of variables for this model.

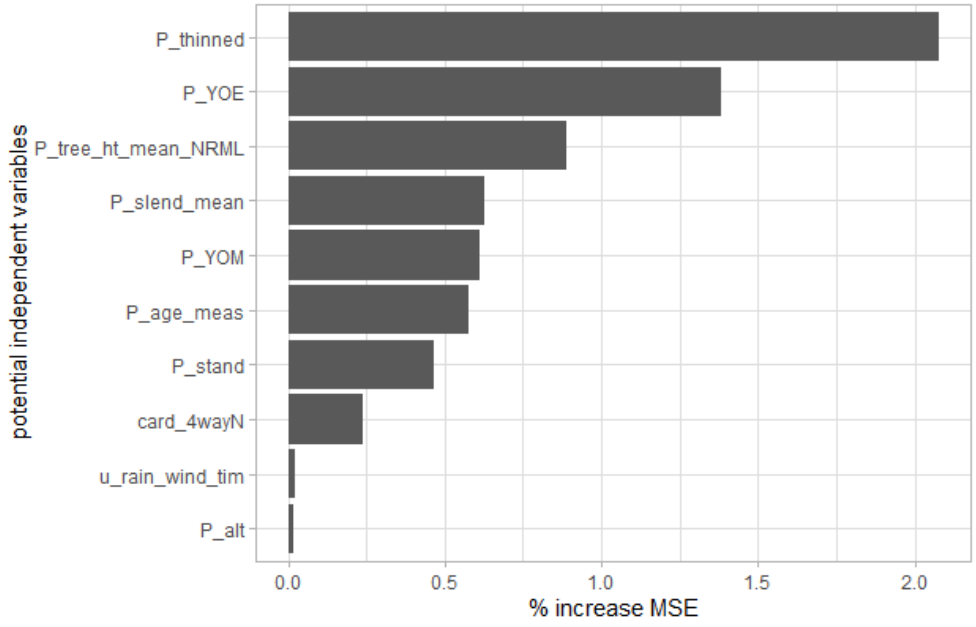

Figure 3-5: relative importance of variables for best random forest model for Douglas-fir P_tree_ht_mean_BRKN.

Examination of the plot of predicted versus actual values (Figure 3-6, below) shows a similar pattern to the corresponding plot (Figure 3-2) for modelling of Douglas-fir *P_tree_ht_mean_BRKN* by regression analysis, with the data again sorting into two bands. Examination of the variable importance measures (not presented here) shows that P_thinned has strongest influence, causing the two-bands effect, as was the case in the corresponding regression analysis (section 3.3.1.2).

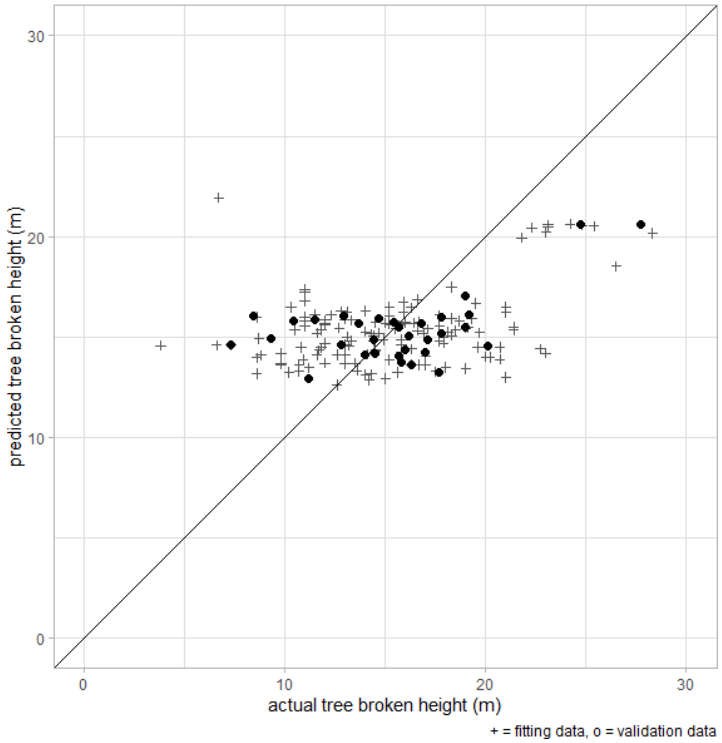

Figure 3-6: Visualising best random forest model for Douglas-fir P_tree_ht_mean_BRKN. Actual and predicted values from validation and fitting data; 1:1 line for reference.

3.4 Species-level results for the proportion of damaged trees

This section details, for each species, the models created to investigate the relationship between the response variables proportion of damaged trees per plot *(Tops_prpn_DAM)* and the explanatory variables.

3.4.1 Proportion of damaged trees per plot modelled with logistic regression

Multivariate logistic regression was used to model proportion of damaged trees per plot *(Tops_prpn_DAM)*, for radiata pine based on all radiata pine plots, for radiata based on only plots where the top status of all trees is known, and for Douglas-fir. The candidate variables and the variables selected for inclusion are shown in Table 3-7, Table 3-8, and Table 3-9. All models benefited from the inclusion of the observation-level random effect (OLRE) *Plot_no*, which illustrates that over-dispersion is an issue for this response variable.

As these are logistic regression models, the model coefficients for continuous variables represent the log of the odds of the change in outcome expected for a change of one unit in that variable. Because these continuous variables are centred by the mean, and standardised by the standard deviation, one unit is one standard deviation. Categorical variable coefficients give the log of the odds of the shift in the model intercept expected for the level of the category. Random intercepts represent the variance in the model intercept arising from membership of that random effect group.

3.4.1.1 Radiata pine, all plots included

The results for radiata pine *Tops_prpn_DAM,* as shown in Table 3-7, below, describe a model with moderate explantory power in the fitting step and poor explanatory power in the validation step. The model fitting R^2 is 0.579, the model validation R^2 is 0.121, the slope for the model fitting bias check 0.429, and the slope for the model validation bias check is 0.198. Note also that in this model all the levels of card_4wayN are significantly different from one another.

Table 3-7: details of best logistic regression model, for radiata pine Tops_prpn_DAM, for all plots.

model type	**multivariate logistic regression with mixed-effects by random intercepts**						
candidate variables	fixed: P_age_meas_cs, P_BA_ha_equiv_cs, P_sph_equiv_cs, P_slend_mean_cs, P_Fk_1_prpn_cs, P_slope_cs, P_alt_cs, card_4wayN, card_4wayNE, card_8way, MPI_100_cs, MPI_200_cs, MPI_500_cs, MPI_1000_cs, MPI_2000_cs, WindSheltS1_cs, WindSheltNE1_cs, P_pruned, P_pru_prpn_cs, P_pru_ht_cs, P_thinned, u_wind_tim_cs, u_rain_cs, u_min_temp_cs, u_mint_rain_cs, u_air_pr_cs, u_rain_wind_tim_cs, P_dbh_mean_NRML_cs, P_tree_ht_mean_NRML_cs, Prpn_LIVE_cs random: Plot_no, P_YOE, P_YOM, P_stand						
model fitting	513 observations, 465 above 0						
fixed effects	variable	coefficient	coef. as % change odds		std. error	p-value	
	(intercept)	0.117	(n/a)		0.212	0.5789	
	card_4wayN – N	-0.209	-19		0.094	0.0258	
	card_4wayN – S	-0.260	-23		0.139	0.0615	
	card_4wayN – W	-0.499	-39		0.119	<0.0001	
	MPI_1000_cs	0.241	27		0.041	<0.0001	
	P_thinned – UT	0.383	46		0.155	0.0137	
	Prpn_LIVE_cs	-0.205	-99		0.035	<0.0001	
random intercepts	group	levels	variance		std. dev.		
	Plot_no (OLRE)	510	0.170		0.413		
	P_YOE	20	0.118		0.343		
	P_YOM	8	0.194		0.440		
fit statistics	R^2, this model	R^2, without mixed effects	mean	MAPE	bias: intercept	bias: slope	dispersion factor
	0.579	0.113	0.460	11.7	0.275	0.429	0.629
autocorrelation of residuals	Moran's I observed	Moran's I expected	p-value		autocorrelated		
	0.0147	-0.0020	0.0008		yes		
model validation	112 observations, 105 above 0						
fit statistics	R^2	mean	MAPE	bias: intercept		bias: slope	
	0.121	0.466	15.9	0.379		0.198	

Plotting (Figure 3-7) the actual and predicted damaged illustrates the poorer fit of the validation data, compared to the fitting data. In general terms, the model fit for both validation and fitting data are imprecise and somewhat biased.

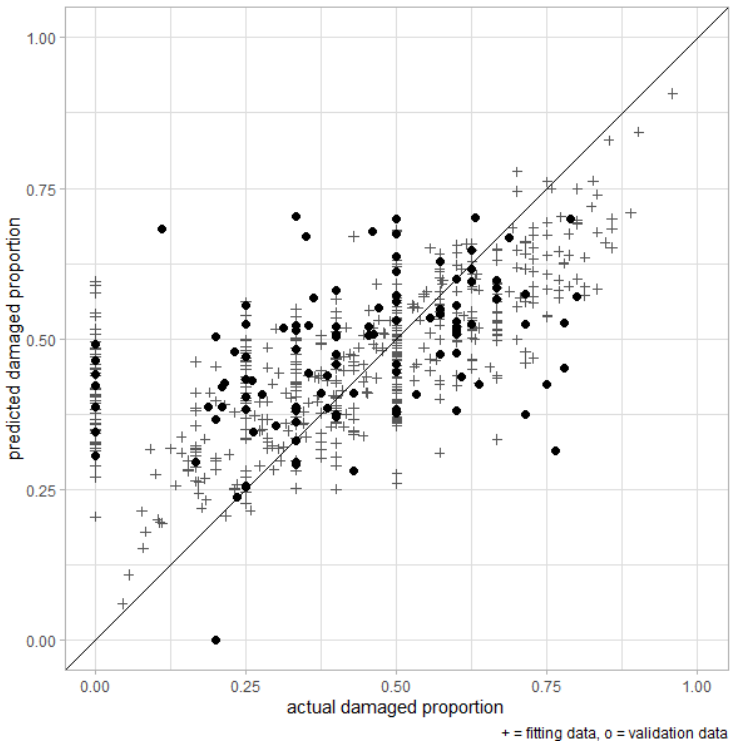

Figure 3-7: Visualising best regression model for radiata pine Tops_prpn_DAM, for all plots. Actual and predicted values from validation and fitting data; 1:1 line for reference.

3.4.1.2 Radiata pine, all plots included, by manual hurdle model

Results for this modelling technique are shown in Table 3-8, below. Modelling step 1) has a poor ability to predict no trees damaged: the apparent error rate (incorrect predictions as a proportion of all predictions) is quite low at 0.088, but the errors are unbalanced, being mostly predictions of some tree damage when none was present. Modelling step 2) has low explanatory power, with an R^2 of 0.179 and a bias check slope of 0.254.

This model yielded somewhat improved fit statistics for the validation data than for the multivariate logistic regression for the same data shown in section 3.4.1.1 (R^2 = 0.179 versus R^2 = 0.121, and bias check slope 0.254 versus 0.198; see also Table 3-7).

Table 3-8: manual hurdle model for radiata pine Tops_prpn_DAM, all plots.

model types	**binary logistic regression and multivariate logistic regression**				
candidate variables	fixed: P_age_meas_cs, P_BA_ha_equiv_cs, P_sph_equiv_cs, P_slend_mean_cs, P_Fk_1_prpn_cs, P_slope_cs, P_alt_cs, card_4wayN, card_4wayNE, card_8way, MPI_100_cs, MPI_200_cs, MPI_500_cs, MPI_1000_cs, MPI_2000_cs, WindSheltS1_cs, WindSheltNE1_cs, P_pruned, P_pru_prpn_cs, P_pru_ht_cs, P_thinned, u_wind_tim_cs, u_rain_cs, u_min_temp_cs, u_mint_rain_cs, u_air_pr_cs, u_rain_wind_tim_cs, P_dbh_mean_NRML_cs, P_tree_ht_mean_NRML_cs, Prpn_LIVE_cs random: Plot_no, P_YOE, P_YOM, P_stand				
step 1 – binary logistic regression, for all plots					
model type	binary logistic regression				
model fitting	509 observations (P_slend_mean has incomplete coverage)				
	variable	coefficient	coef. as % change odds	std. error	p-value
	intercept	2.618	(n/a)	0.202	<0.0001
	P_age_meas_cs	0.314	37	0.176	0.0739
	P_sph_equiv_cs	1.169	222	0.258	<0.0001
	P_slend_mean_cs	-0.555	-43	0.184	0.0026
	MPI_500_cs	0.487	62	0.164	0.0031
model validation					
fit statistics	apparent error rate	confusion matrix	none damaged predicted		some damaged predicted
	0.088	none damaged actual	3 *(correct)*		45 (*incorrect*)
		some damaged actual	0 (*incorrect*)		461 (*correct*)

step 2 – multivariate logistic regression, only plots with some trees damaged					
model type	logistic regression with mixed-effects				
model fitting	461 observations (P_dbh_mean_NRML_cs has incomplete coverage)				
fixed effects	variable	coefficient	coef. as % change odds	std. error	p-value
	(intercept)	0.092	(n/a)	0.190	0.6310
	P_sph_equiv_cs	0.241	27	0.045	<0.0001
	P_alt_cs	0.198	21	0.051	0.0001
	card_4wayN – N	-0.331	-28	0.090	0.0002
	card_4wayN – S	-0.364	-30	0.137	0.0078
	card_4wayN – W	-0.628	-47	0.115	<0.0001
	MPI_500_cs	0.288	33	0.041	<0.0001
	u_wind_tim_cs	-0.303	-26	0.087	0.0005
	P_dbh_mean_NRML_cs	0.205	22	0.065	0.0015
	Prpn_LIVE_cs	-0.120	-11	0.034	0.0004

random intercepts	group	levels	variance	std. dev.
	Plot_no (OLRE)	461	0.125	0.354
	P_YOE	20	0.505	0.711

fit statistics	R^2, this model	R^2, without mixed effects	mean	MAPE	bias: intercept	bias: slope	dispersion factor
	0.634	0.135	0.470	9.4	0.218	0.537	0.557

autocorrelation of residuals	Moran's I observed	Moran's I expected	p-value	autocorrelated
	0.0021	-0.0022	0.4482	no

model validation (for combined effects of above models on validation data)					
fit statistics	R^2	mean	MAPE	bias: intercept	bias: slope
	0.179	0.478	14.6	0.369	0.254

Plotting the actual versus predicted values for the validation set and fitting sets together (Figure 3-8) illustrates the results of using this model. The 45 values that fell into the category *no actual damage/some predicted damage* in the AER section of Table 3-8, above, are clearly visible in the plot as a vertical band at zero actual damage predicted.

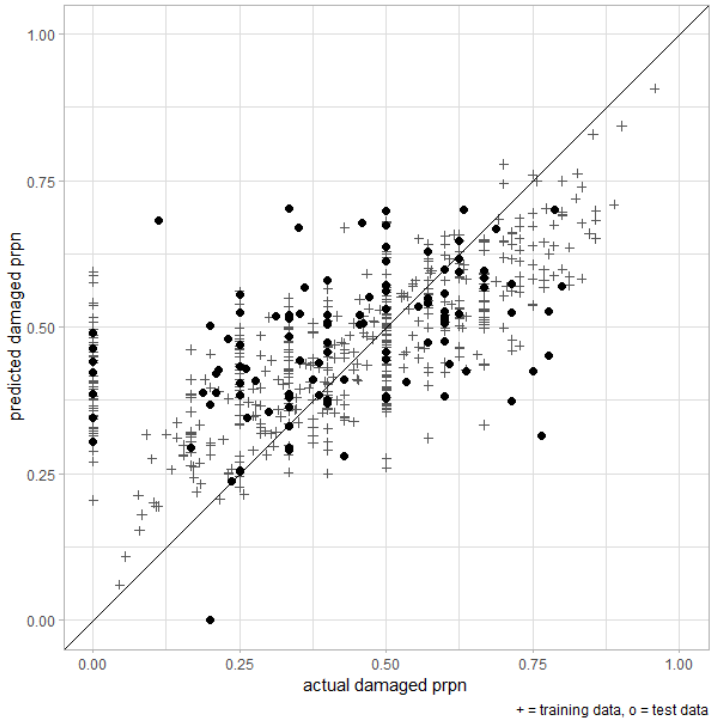

Figure 3-8: Validating manual hurdle model for radiata pine Tops_prpn_DAM. Actual and predicted values from validation and fitting data; 1:1 line for reference.

3.4.1.3 Radiata pine, only plots with full tops assessment

If only those radiata pine plots that had all tops assessed are included, a model that has moderate R^2 (0.808 fitting, 0.545 validation) and moderate bias slope (0.646 fitting, 0.385 validation) may be created. The influence of aspect is weak, with only card_4wayN: west being significantly different the other levels of the category. The variable *P_YOM* is not a candidate for a random effect in this case, because the vast majority of the plots that had all tops assessed were measured in a single year.

Table 3-9: details of best logistic regression model, for radiata pine Tops_prpn_DAM, for plots with all tops assessed only.

model type	**multivariate logistic regression with mixed-effects**						
candidate variables	fixed: P_age_meas_cs, P_BA_ha_equiv_cs, P_sph_equiv_cs, P_slend_mean_cs, P_Fk_1_prpn_cs, P_slope_cs, P_alt_cs, card_4wayN, card_4wayNE, card_8way, MPI_100_cs, MPI_200_cs, MPI_500_cs, MPI_1000_cs, MPI_2000_cs, WindSheltS1_cs, WindSheltNE1_cs, P_pruned, P_pru_prpn_cs, P_pru_ht_cs, P_thinned, u_wind_tim_cs, u_rain_cs, u_min_temp_cs, u_mint_rain_cs, u_air_pr_cs, u_rain_wind_tim_cs, P_dbh_mean_NRML_cs, P_tree_ht_mean_NRML_cs, Prpn_LIVE_cs random: Plot_no, P_YOE, P_stand						
model fitting	166 observations, 162 above 0						
fixed effects	variable	coefficient	coef. as % change odds		std. error	p-value	
	intercept	-0.393	(n/a)		0.090	<0.0001	
	P_sph_equiv_cs	0.302	35		0.071	<0.0001	
	card_4wayNE – NW	-0.201	-18		0.120	0.0938	
	card_4wayNE – SE	0.152	16		0.208	0.4644	
	card_4wayNE – SW	-0.314	-27		0.165	0.0572	
	MPI_200_cs	0.302	35		0.056	<0.0001	
	P_dbh_mean_NRML_cs	0.235	26		0.071	<0.0001	
	Prpn_LIVE_cs	-0.274	-23		0.056	<0.0001	
random intercepts	group	levels			variance	std. dev.	
	Plot_no (OLRE)	166			0.142	0.377	
	P_stand	54			0.046	0.214	
fit statistics	R^2, this model	R^2, without mixed effects	mean	MAPE	bias: intercept	bias: slope	dispersion factor
	0.799	0.387	0.387	6.8	0.139	0.641	0.607
autocorrelation of residuals	Moran's I observed	Moran's I expected			p-value	autocorrelated	
	0.0061	-0.0061			0.9999	no	
model validation	33 observations, 32 above 0						
fit statistics	R^2	mean	MAPE	bias: intercept		bias: slope	
	0.416	0.373	13.9	0.253		0.311	

A plot (Figure 3-9, below) of the actual and predicted damaged proportions illustrates the explanatory power of this model:

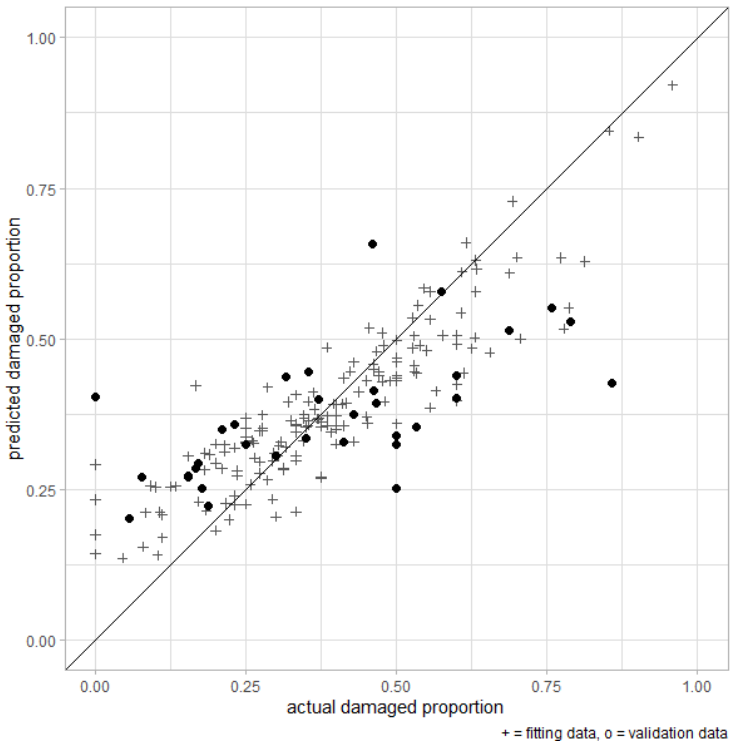

Figure 3-9: Visualising best regression model for radiata pine Tops_prpn_DAM, only for plots with all tops assessed. Actual and predicted values from validation and fitting data; 1:1 line for reference.

3.4.1.4 Douglas-fir

Table 3-10, below, describes a logistic regression model for Douglas-fir *Tops_prpn_DAM* that has overall low explanatory power: the model fitting R^2 is 0.401, the model validation R^2 is 0.259, the slope for the model fitting bias check 0.340, and the slope for the model validation bias check is 0.321. As was the case for radiata pine (see section 3.4.1.1), the model has difficulty accurately predicting true zeros, and is showing data clustering at 0.25, 0.33 and 0.5, as is visible in Figure 3-10.

Table 3-10: details of best logistic regression model, for Douglas-fir Tops_prpn_DAM.

model type	**random forest with conditional inference trees**					
candidate variables	fixed: P_age_meas_cs, P_BA_ha_equiv_cs, P_sph_equiv_cs, P_slend_mean_cs, P_Fk_1_prpn_cs, P_slope_cs, P_alt_cs, card_4wayN, card_4wayNE, card_8way, MPI_100_cs, MPI_200_cs, MPI_500_cs, MPI_1000_cs, MPI_2000_cs, WindSheltS1_cs, WindSheltNE1_cs, P_thinned, u_wind_tim_cs, u_rain_cs, u_min_temp_cs, u_mint_rain_cs, u_air_pr_cs, u_rain_wind_tim_cs, P_dbh_mean_NRML_cs, P_tree_ht_mean_NRML_cs, Prpn_LIVE_cs (no prune vars) random: Plot_no, P_YOE, P_stand (P_YOM excluded as <5 levels)					
model type	multivariate logistic regression with mixed-effects					
model fitting	263 observations, 140 above 0					
fixed effects	variable	coefficient	coef. as % change odds	std. error	p-value	
	(intercept)	-0.875	(n/a)	0.239	0.0003	
	P_BA_ha_equiv_cs	-0.168	-15	0.085	0.0477	
	card_4wayNE – NW	-0.581	-44	0.273	0.0333	
	card_4wayNE – SE	-0.269	-24	0.218	0.2173	
	card_4wayNE – SW	-0.795	-55	0.245	0.0012	
	Prpn_LIVE_cs	-0.223	-20	0.083	0.0073	
random intercepts	group	levels	variance		std. dev.	
	Plot_no (OLRE)	263	0.089		0.298	
	P_stand	18	0.347		0.589	
fit statistics	R^2, this model	R^2, without mixed effects	mean	MAPE	bias: intercept	bias: slope	dispersion factor
	0.401	0.114	0.254	15.2	0.184	0.340	0.757
autocorrelation of residuals	Moran's I observed	Moran's I expected			p-value	autocorrelated	
	-0.0034	-0.0038			0.969	no	
model validation	54 observations, 36 above 0					
fit statistics	R^2	mean	MAPE	bias: intercept	bias: slope	
	0.283	0.275	16.4	0.186	0.321	

Plotting the actual versus predicted values for the validation set and fitting sets together (Figure 3-10) illustrates the highly imprecise results of using this model.

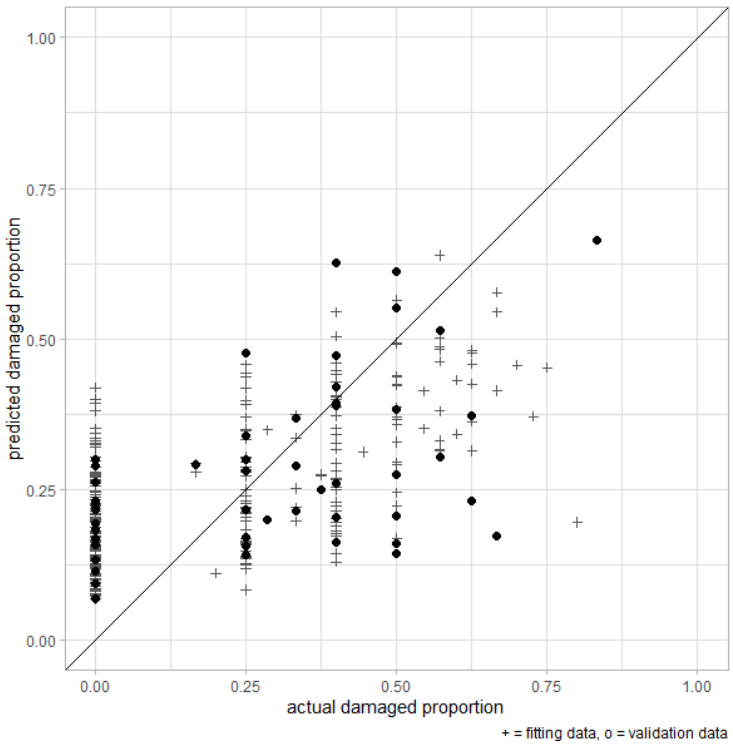

Figure 3-10: Visualising best regression model for Douglas-fir Tops_prpn_DAM. Actual and predicted values from validation and fitting data; 1:1 line for reference.

3.4.2 Proportion of damaged trees per plot modelled with Random Forests

Three alternative random forest models were created for plot proportion of damaged trees (*Tops_prpn_DAM),* for each of radiata pine all plots, radiata pine plots with all tops assessed, and Douglas-fir all plots. The three alternatives were all variables included, the top ten variables by explanatory power, and variables analogous to those used in the best corresponding regression model. Results of the best model are shown below, and the other results may be found in Appendix 6.7. As random forests do not have model coefficients, a plot of the relative importance of variables has been included.

3.4.2.1 Radiata pine, all plots included

The best random forest model of *Tops_prpn_DAM* for radiata pine, with all plots included, has low explanatory power. The model has these has these fit statistics: model fitting R^2 0.267, model validation R^2 0.259, slope for the model fitting bias check 0.243, and slope for the model validation bias check is 0.224. These fit statistics, shown in Table 3-11 below, are worse for fitting but better than for validation data, compared to the corresponding regression model for radiata pine *Tops_prpn_DAM* with all tops included (section 3.4.1.1). This may reflect the absence of any analogue of the observation-level random effect that was used in the corresponding regression.

Table 3-11: details of random forest models of radiata pine Tops_prpn_DAM, including model fit statistics and identification of best model, including all plots.

model type	**random forest with conditional inference trees**
candidate variables	P_age_meas, P_BA_ha_equiv, P_sph_equiv, P_slend_mean, P_Fk_1_prpn, P_slope, P_alt, card_4wayN, card_4wayNE, card_8way, MPI_100, MPI_200, MPI_500, MPI_1000, MPI_2000, WindSheltS1, WindSheltNE1, P_pruned, P_pru_prpn, P_pru_ht, P_thinned, u_wind_tim, u_rain, u_min_temp, u_mint_rain, u_air_pr, u_rain_wind_tim, P_dbh_mean_NRML, P_tree_ht_mean_NRML, Prpn_LIVE, Plot_no, P_YOE, P_YOM, P_stand

model fitting	513 observations, 465 above 0						
choice of variables	variables in model	variables at each split	R^2	mean	MAPE	bias: intercept	bias: slope
top ten variables by explanatory power	P_YOM, P_pru_prpn, P_YOE, MPI_200, card_4wayN, P_slope, MPI_100, P_sph_equiv, P_stand, P_alt	4	0.267	0.438	15.5	0.333	0.243

fit statistics best model	Moran's I observed	Moran's I expected	p-value	autocorrelated
	-0.0009	-0.0019	0.835	no

model validation	112 observations, 105 above 0				
fit statistics	R^2	mean	MAPE	bias: intercept	bias: slope
best model	0.259	0.434	15.2	0.339	0.224

Figure 3-11, below, shows the relative importance of variables for this model.

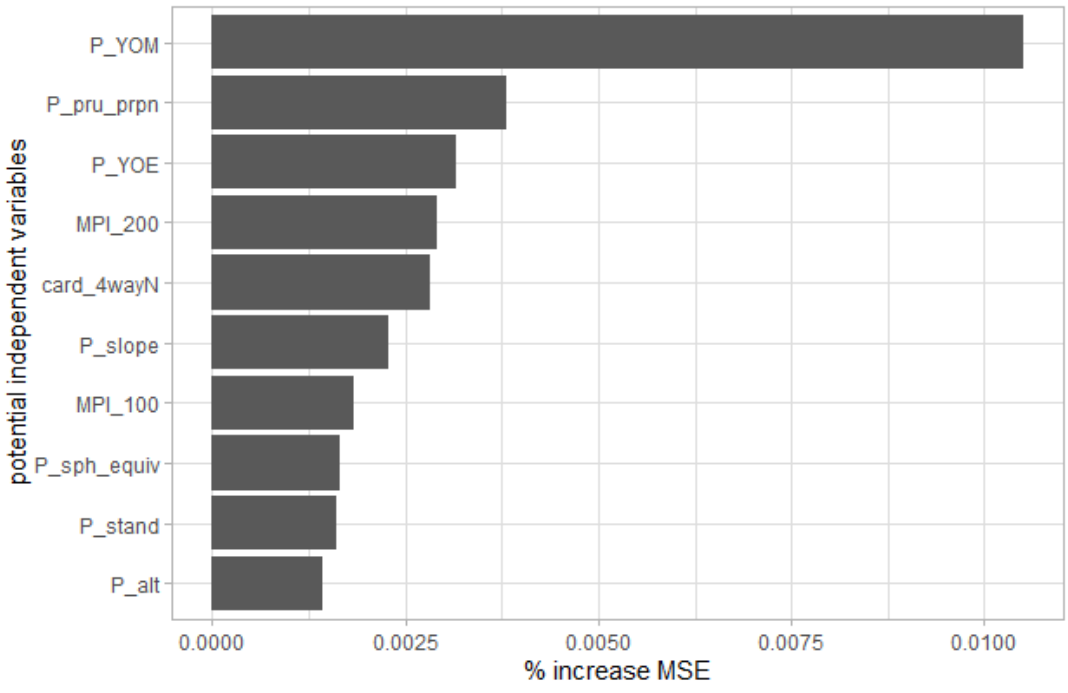

Figure 3-12: relative importance of variables for best random forest model for radiata pine Tops_prpn_DAM, including all plots.

Plotting the actual versus predicted values for the validation set and fitting sets together (Figure 3-13) illustrates the results of using this model. Comparison with Figure 3-7, from the corresponding regression analysis, shows the relatively greater bias in these model outcomes.

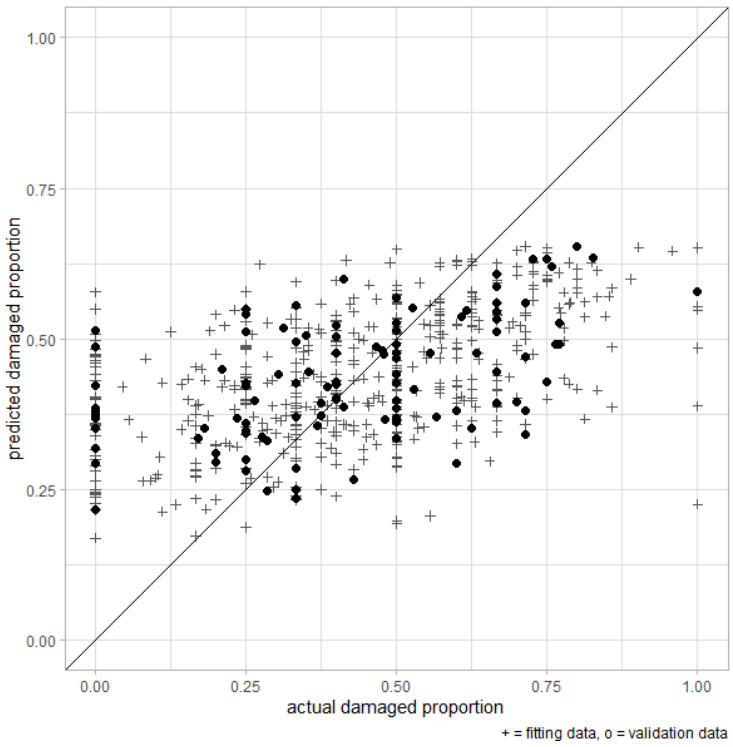

Figure 3-13: Visualising best random forest model for radiata pine Tops_prpn_DAM, for all plots. Actual and predicted values from validation and fitting data; 1:1 line for reference.

3.4.2.2 Radiata pine, only plots with full tops assessment

The best random forest model of *Tops_prpn_DAM* for radiata pine, including only plots where all tops of trees were assessed, has low explanatory power. The model has these fit statistics: model fitting R^2 0.226, the model validation R^2 0.325, slope for the model fitting bias check 0.262, and slope for the model validation bias check is 0.261. Full results are shown in Table 3-12, below.

Restriction of this random forest to the subset of plots with all tops assessed has made little difference to the explantory power, compared to the corresponding random forest with all plots included (section 3.4.2.1; for easy comparison of fit statistics, see Table 3-2). This is unlike the regression models, where the model formed from the subset of plots with all tops assessed (section 3.4.1.3) had stronger explanatory power than the model with all plots included (section 3.4.1.1).

Table 3-12: details of random forest models of radiata pine Tops_prpn_DAM, including model fit statistics and identification of best model, for plots with all tops assessed only.

model type	**random forest with conditional inference trees**
candidate variables	P_age_meas, P_BA_ha_equiv, P_sph_equiv, P_slend_mean, P_Fk_1_prpn, P_slope, P_alt, card_4wayN, card_4wayNE, card_8way, MPI_100, MPI_200, MPI_500, MPI_1000, MPI_2000, WindSheltS1, WindSheltNE1, P_pruned, P_pru_prpn, P_pru_ht, P_thinned, u_wind_tim, u_rain, u_min_temp, u_mint_rain, u_air_pr, u_rain_wind_tim, P_dbh_mean_NRML, P_tree_ht_mean_NRML, Prpn_LIVE, Plot_no, P_YOE, P_stand

model fitting	166 observations, 162 above 0						
choice of variables	variables in model	variables at each split	R^2	mean	MAPE	bias: intercept	bias: slope
top ten variables by explanatory power	Prpn_LIVE, MPI_1000, MPI_500, P_sph_equiv, P_slope, MPI_2000, MPI_200, MPI_100, P_pru_prpn, P_YOE	4	0.226	0.390	13.0	0.292	0.262

fit statistics best model	Moran's I observed	Moran's I expected	p-value	autocorrelated
	0.0078	0.0130	0.2871	no

model validation	33 observations, 32 above 0				
fit statistics best model	R^2	mean	MAPE	bias: intercept	bias: slope
	0.325	0.392	14.7	0.292	0.261

Figure 3-14, below, shows the relative importance of variables for this model.

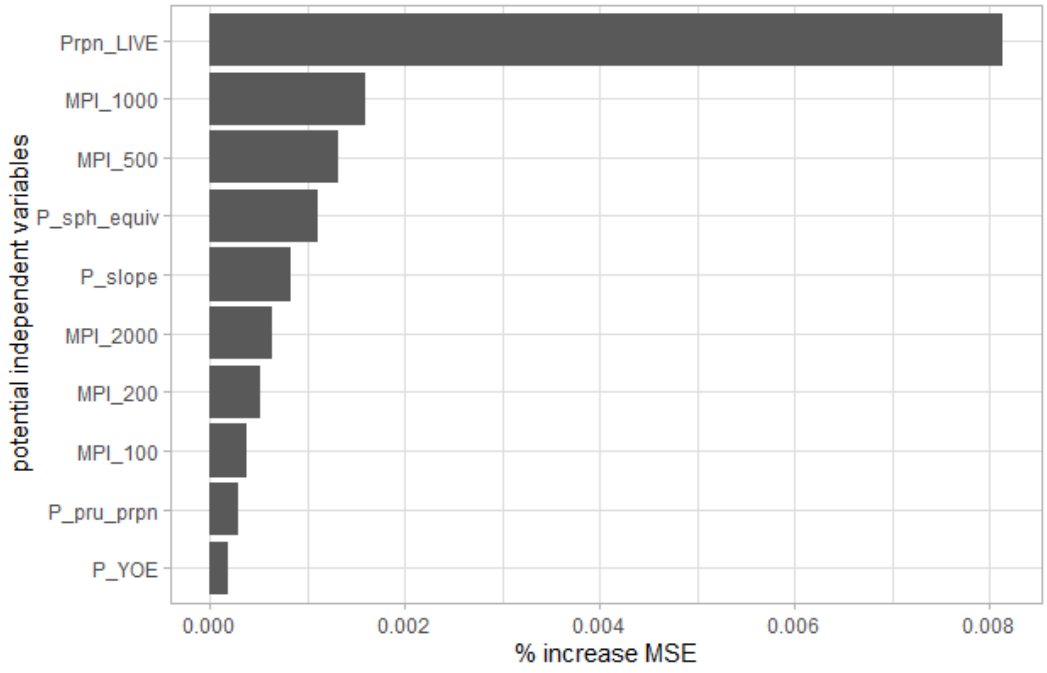

Figure 3-15: relative importance of variables for best random forest model for radiata pine Tops_prpn_DAM, including only plots with all tops assessed.

Examination of the plot of predicted versus actual values (Figure 3-16, below) compared with the corresponding plot (Figure 3-9) for modelling of radiata pine *Tops_prpn_DAM* by regression analysis reinforces what the table above shows: the corresponding regression model has lower bias and higher precision.

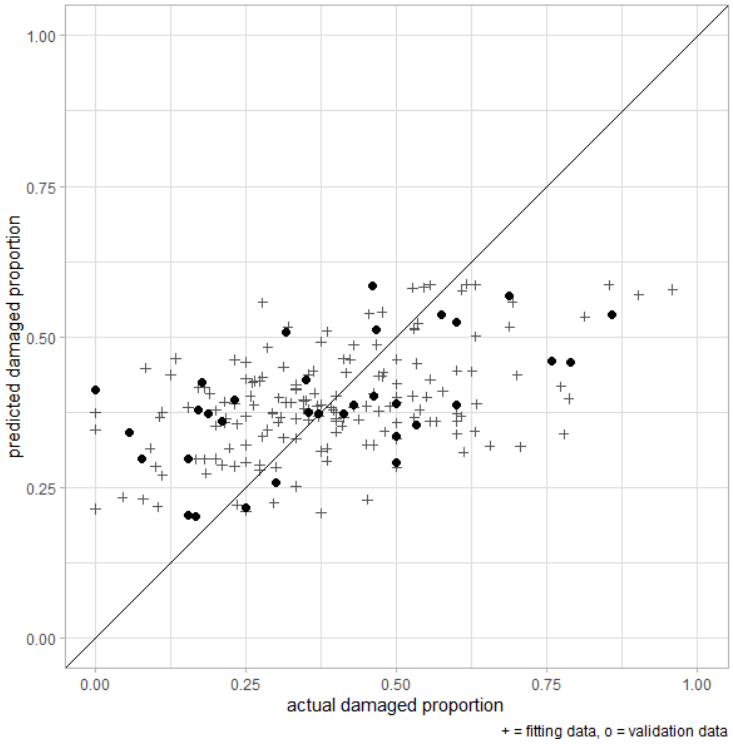

Figure 3-16: Visualising best random forest model for radiata pine Tops_prpn_DAM, for plots with all tops asssessed only. Actual and predicted values from validation and fitting data; 1:1 line for reference.

3.4.2.3 Douglas-fir

The best random forest model of proportion damaged per plot (*Tops_prpn_DAM*) for Douglas-fir is presented in Table 3-13 and Figure 3-27 below. The model has these fit statistics: model fitting R^2 0.401, the model validation R^2 0.191, slope of the model fitting bias check 0.196, the slope for the model validation bias check 0.215.

Table 3-13: details of random forest models of Douglas-fir Tops_prpn_DAM, including model fit statistics and identification of best model.

model type	**random forest with conditional inference trees**						
candidate variables	P_age_meas, P_BA_ha_equiv, P_sph_equiv, P_slend_mean, P_Fk_1_prpn, P_slope, P_alt, card_4wayN, card_4wayNE, card_8way, MPI_100, MPI_200, MPI_500, MPI_1000, MPI_2000, WindSheltS1, WindSheltNE1, P_thinned, u_wind_tim, u_rain, u_min_temp, u_mint_rain, u_air_pr, u_rain_wind_tim, P_dbh_mean_NRML, P_tree_ht_mean_NRML, Prpn_LIVE, Plot_no, P_YOE, P_YOM, P_stand (no prune vars)						
model fitting	263 observations, 140 above 0						
choice of variables	variables in model	variables at each split	R^2	mean	MAPE	bias: intercept	bias: slope
top ten variables by explanatory power	card_8way, P_stand, Estab_sph, card_4wayN, MPI_2000, P_YOE, MPI_500, Prpn_LIVE, WindSheltNE1, card_4wayNE	4	0.191	0.228	16.4	0.174	0.196

fit statistics best model	Moran's I observed	Moran's I expected	p-value	autocorrelated
	0.0090	-0.039	0.2135	yes

model validation	54 observations, 36 above 0				
fit statistics best model	R^2	mean	MAPE	bias: intercept	bias: slope
	0.324	0.387	18.2	0.175	0.215

Figure 3-17, below, shows the relative importance of variables for this model.

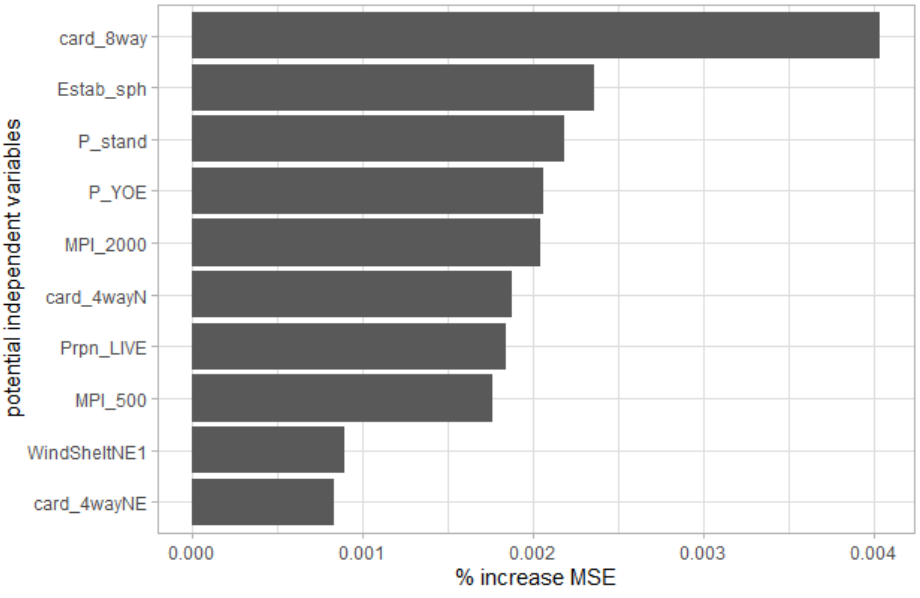

Figure 3-18: relative importance of variables for best random forest model for Douglas-fir Tops_prpn_DAM.

Plotting the actual versus predicted values for the validation set and fitting sets together (Figure 3-1) illustrates the results of using this model.

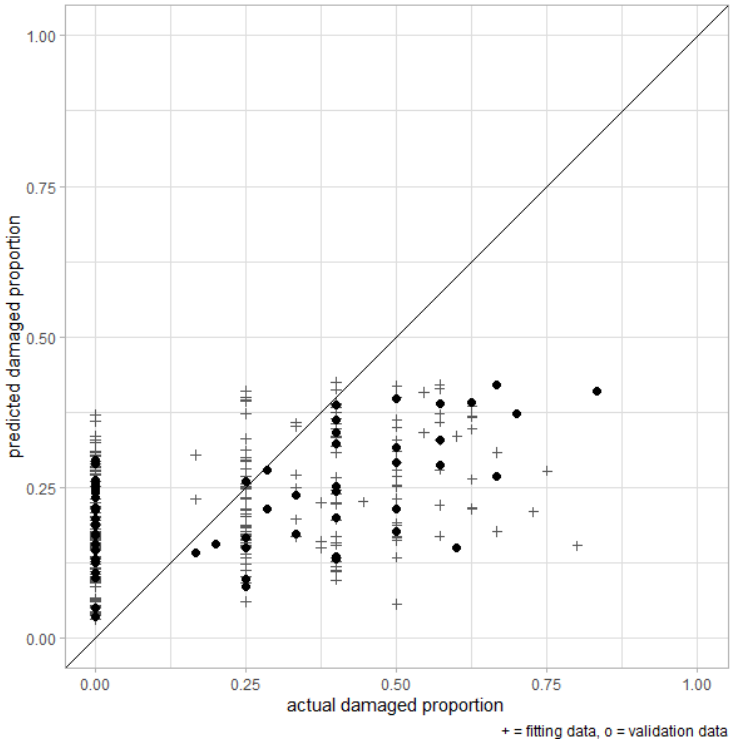

Figure 3-19: Visualising best random forest model for Douglas-fir Tops_prpn_DAM. Actual and predicted values from validation and fitting data; 1:1 line for reference.

3.5 Species-level results for the proportion of live trees

This section details, for each species, the models created to investigate the relationship between the response variables proportion of live trees per plot *(Prpn_LIVE)* and the explanatory variables. Although *Prpn_LIVE* is not as direct a metric of tree damage as *P_mean_ht_BRKN* or *Tops_prpn_DAM*, increased plot damage has a discernible negative correlation with the proportion of trees alive: - 0.11 for radiata pine and - 0.17 for Douglas-fir.

3.5.1 Proportion of live trees per plot modelled with logistic regression

Logistic regression was used to model the proportion of live trees per plot (*Prpn_LIVE*) for radiata pine and Douglas-fir. Both models benefited from the inclusion of the observation-level random effect (OLRE) *Plot_no*, which illustrates that over-dispersion is an issue for this response variable. Because these models are also logistic regression models, the same interpretation of model coefficients applies as for section 3.4.1.

3.5.1.1 Radiata pine

The results for radiata pine proportion of live trees per plot (*Prpn_LIVE*), as shown in Table 3-14 below, describe a model with that has much stronger explanatory power in the model fitting step than the model validation step. The model has these fit statistics: model fitting R^2 is 0.807, model validation R^2 0.471, slope for the model fitting bias check 0.386, and slope for the model validation bias check is 0.079. This is possibly attributable to the heavy reliance of the fitted model on the OLRE, which manifests in the very different R^2 for this model and the related variant without the OLRE (0.807 versus 0.117). The reasons why OLREs are not particularly helpful for predicting new data are examined in section 4.5.1.

Table 3-14: details of best logistic regression model, Prpn_LIVE.

model type	**multivariate logistic regression with mixed-effects**
candidate variables	fixed: P_age_meas_cs , P_BA_ha_equiv_cs, P_sph_equiv_cs, P_slend_mean_cs, P_Fk_1_prpn, P_slope_cs, P_alt_cs, card_4wayN, card_4wayNE, card_8way,MPI_100_cs, MPI_200_cs, MPI_500_cs, MPI_1000_cs, MPI_2000_cs, WindSheltS1_cs, WindSheltNE1_cs, P_pruned, P_pru_prpn_cs, P_pru_ht_cs, P_thinned, Estab_sph_cs, Sph_drop_cs, u_wind_tim_cs, u_rain_cs, u_min_temp_cs, u_mint_rain_cs, u_air_pr_cs, u_rain_wind_tim_cs, P_dbh_mean_NRML_cs, P_tree_ht_mean_NRML_cs, Tops_prpn_DAM_cs random: Plot_no, P_YOE, P_YOM, P_stand

model fitting	513 observations, 106 less than 1				
fixed effects	variable	coefficient	coef. as % change odds	std. error	p-value
	intercept	5.040	(n/a)	0.237	<0.0001
	P_sph_equiv_cs	-0.366	-31	0.108	0.0006
	P_alt_cs	0.290	34	0.107	0.0067
	card_4wayNE – NW	-0.387	-32	0.236	0.1007
	card_4wayNE – SE	1.088	197	0.448	0.0151
	card_4wayNE – SW	-0.505	-40	0.341	0.1390
	P_pru_prpn_cs	0.180	20	0.108	0.0952
	Estab_sph_cs	0.325	38	0.122	0.0075
	u_wind_tim_cs	-0.736	-52	0.143	<0.0001
	Tops_prpn_DAM_cs	-0.287	-25	0.116	0.0135

random intercepts	group	levels	variance	std. dev.
	Plot_no (OLRE)	513	0.960	0.980

fit statistics	R^2, this model	R^2, without mixed effects	mean	MAPE	bias: intercept	bias: slope	dispersion factor
	0.750	0.117	0.986	1.4	0.523	0.471	0.484

autocorrelation of residuals	Moran's I observed	Moran's I expected	p-value	autocorrelated
	-0.0070	-0.0020	0.3066	no

model validation	112 observations, 29 less than 1				
fit statistics	R^2	mean	MAPE	bias: intercept	bias: slope
	0.117	0.991	2.01	0.913	0.079

Plotting the actual and predicted values for the fitting and validation steps data (Figure 3-20) illustrates how this model has better explanatory power for the fitting than the validation data, but both data types have biased and imprecise predictions. The plot also shows that data with *Prpn_LIVE* = 1 are problematic to model.

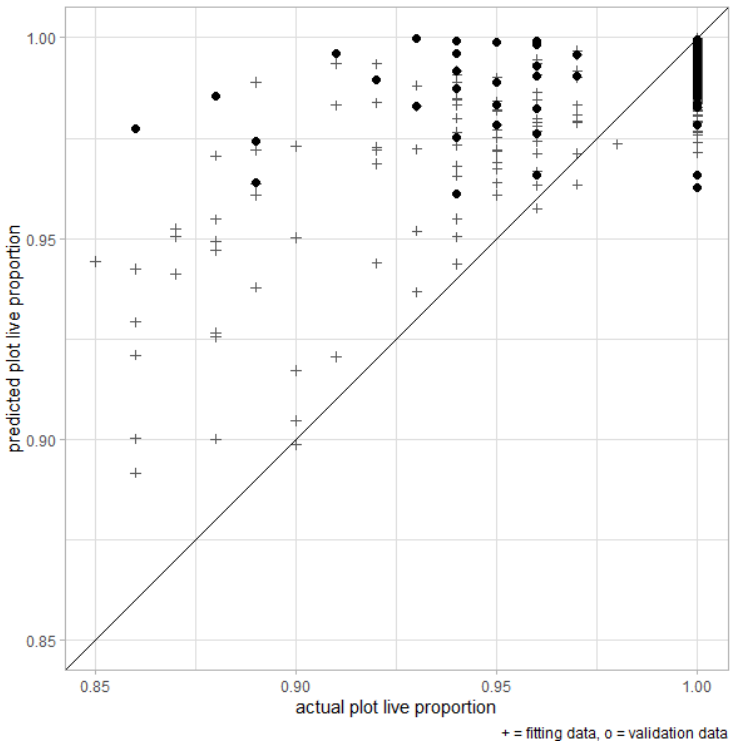

Figure 3-20: Testing best regression for radiata pine Prpn_LIVE. Actual and predicted values from validation and fitting data; 1:1 line for reference.

3.5.1.2 Douglas-fir

The results for Douglas-fir proportion of live trees per plot (*Prpn_LIVE*), as shown in Table 3-15 below, describe a model with low explantory power in the fitting step and very low explantory power in the validation step. The model fitting R^2 is 0.300, the model validation R^2 is 0.114, the slope for the model fitting bias check 0.219, and the slope for the model validation bias check is 0.104. Overall, barely any of the variation in the proportion of live trees is explained by the model fitted.

Table 3-15: details of best random forest model for Douglas-fir Prpn_LIVE.

model type	**random forest with conditional inference trees**					
candidate variables	fixed: P_age_meas_cs , P_BA_ha_equiv_cs, P_sph_equiv_cs, P_slend_mean_cs, P_Fk_1_prpn, P_slope_cs, P_alt_cs, card_4wayN, card_4wayNE, card_8way,MPI_100_cs, MPI_200_cs, MPI_500_cs, MPI_1000_cs, MPI_2000_cs, WindSheltS1_cs, WindSheltNE1_cs, P_thinned, Estab_sph_cs, Sph_drop_cs, u_wind_tim_cs, u_rain_cs, u_min_temp_cs, u_mint_rain_cs, u_air_pr_cs, u_rain_wind_tim_cs, P_dbh_mean_NRML_cs, P_tree_ht_mean_NRML_cs, Tops_prpn_DAM_cs (no prune vars) random: Plot_no, P_YOE, P_stand (P_YOM excluded as <5 levels)					

model type	multivariate logistic regression with mixed-effects					
model fitting	263 observations, 103 less than 1					
fixed effects	variable	coefficient	coef. as % change odds	std. error	p-value	
	intercept	3.691	(n/a)	0.120	<0.0001	
	P_sph_equiv_cs	-0.345	-29	0.093	0.0002	
	P_tree_ht_mean_NRML_cs	-0.426	-35	0.092	<0.0001	
	Tops_prpn_DAM_cs	-0.299	-26	0.085	0.0005	

random intercepts	group	levels	variance	std. dev.
	Plot_no (OLRE)	261	0.095	0.308

fit statistics	R^2, this model	R^2, without mixed effects	mean	MAPE	bias: intercept	bias: slope	dispersion factor
	0.252	0.093	0.972	2.9	0.778	0.200	0.889

autocorrelation of residuals	Moran's I observed	Moran's I expected	p-value	autocorrelated
	0.0083	-0.0038	0.656	no

model validation	54 observations, 23 less than 1				
fit statistics	R^2	mean	MAPE	bias: intercept	bias: slope
	0.114	0.967	4.6	0.866	0.105

Plotting (Figure 3-21) the actual and predicted values for the fitting and validation steps data further underlines the inadequate fit for this model:

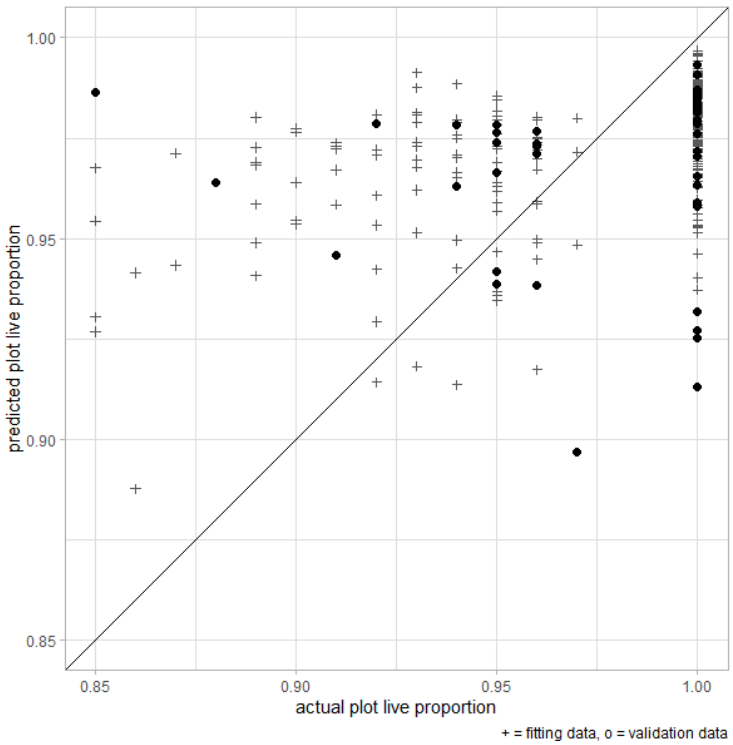

Figure 3-21: Testing best regression for Douglas-fir Prpn_LIVE. Actual and predicted values from validation and fitting data; 1:1 line for reference.

3.5.2 Proportion of live trees per plot modelled with Random Forests

Three alternative random forest models were created for plot proportion of live trees (*Prpn_LIVE),* for each of radiata pine and Douglas-fir. The three alternatives were all variables included, the top ten variables by explanatory power, and variables analogous to those used in the best corresponding regression model. Results of the best model are shown below, and the other results may be found in Appendix 6.7. As random forests do not have model coefficients, a plot of the relative importance of variables has been included.

3.5.2.1 Radiata pine

The best random forest model of pine proportion of live tree per plot (*Prpn_LIVE*) for radiata pine has very low explantory power. The model has these fit statistics, as shown in Table 3-16 below: model fitting R^2 0.184, model validation R^2 0.226, slope for the model fitting bias check is 0.182, and slope for the model validation bias check is 0.146. These fit statistics are worse in fitting but better in validation than for this corresponding regression model for radiata pine *Prpn_LIVE* (section 3.5.1.1), possibly because there is no analogue here for the observation-level random effect that is so influential in the regression model for radiata pine *Prpn_LIVE*.

Table 3-16: details of random forest models of radiata pine Prpn_LIVE, including model fit statistics and identification of best model.

model type	**random forest with conditional inference trees**
candidate variables	P_age_meas , P_BA_ha_equiv, P_sph_equiv, P_slend_mean, P_Fk_1_prpn, P_slope, P_alt, card_4wayN, card_4wayNE, card_8way,MPI_100, MPI_200, MPI_500, MPI_1000, MPI_2000, WindSheltS1, WindSheltNE1, P_pruned, P_pru_prpn, P_pru_ht, P_thinned, Estab_sph, Sph_drop, u_wind_tim, u_rain, u_min_temp, u_mint_rain, u_air_pr, u_rain_wind_tim, P_dbh_mean_NRML, P_tree_ht_mean_NRML, Tops_prpn_DAM, Plot_no, P_YOE, P_YOM, P_stand

model fitting	513 observations, 106 less than 1						
choice of variables	variables in model	variables at each split	R^2	mean	MAPE	bias: intercept	bias: slope
top ten variables by explanatory power	P_YOM, Tops_prpn_DAM, P_stand, P_age_meas, u_wind_tim, P_pruned, u_air_pr, P_pru_prpn, u_rain	4	0.184	0.983	2.1	0.805	0.182

fit statistics best model	Moran's I observed	Moran's I expected	p-value	autocorrelated
	-0.0100	-0.0019	0.1034	no

model validation	112 observations, 29 less than 1				
fit statistics best model	R^2	mean	MAPE	bias: intercept	bias: slope
	0.226	0.982	2.0	0.739	0.247

Figure 3-22, below, shows the relative importance of variables for this model.

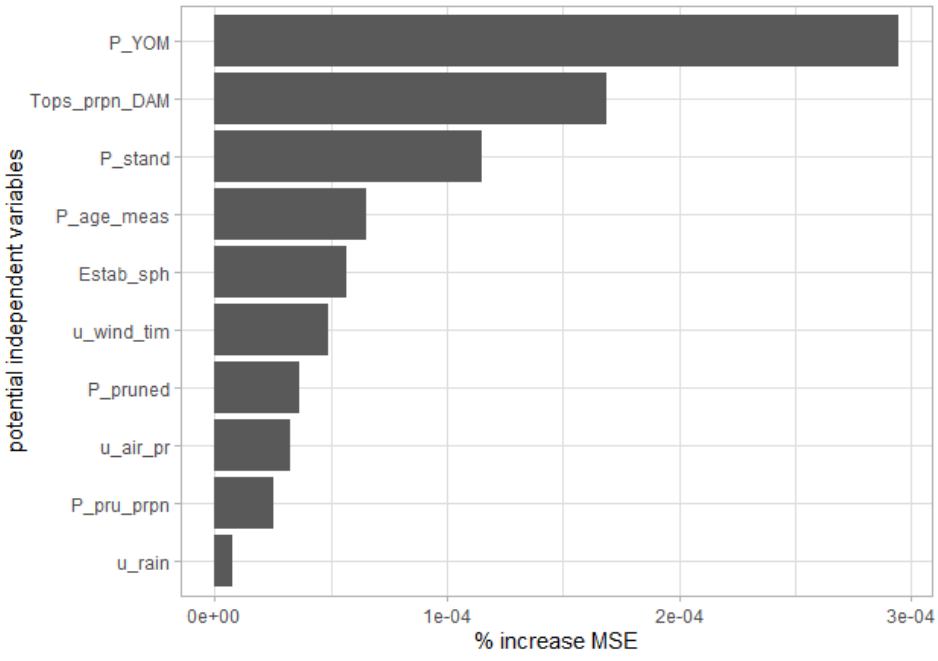

Figure 3-23: relative importance of variables for best random forest model for radiata pine Prpn_LIVE.

Plotting the actual versus predicted values for the validation set and fitting sets together (Figure 3-24) illustrates the results of using this model: a very poor fit, and difficulties predicting when *Prpn_LIVE* equals one.

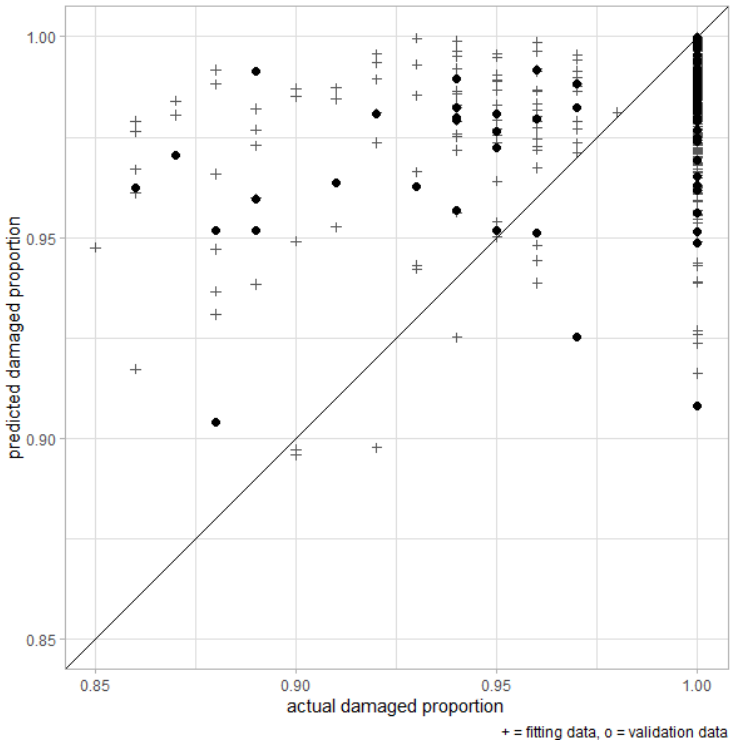

Figure 3-24: Visualising best random forest model for radiata pine Prpn_LIVE. Actual and predicted values from validation and fitting data; 1:1 line for reference.

3.5.2.2 Douglas-fir

The best random forest model for Douglas-fir proportion of live trees per plot (*Prpn_LIVE*) has very low (near-zero) explantory power. The model fit statistics, shown in Table 3-17, are: model fitting R^2 0.097, model validation R^2 0.025, slope for the model fitting bias check 0.108, and slope for the model validation bias check is 0.041.

Table 3-17: details of random forest models of Douglas-fir proportion of live trees per plot (Prpn_LIVE), including model fit statistics and identification of best model.

model type	**random forest with conditional inference trees**
candidate variables	P_age_meas , P_BA_ha_equiv, P_sph_equiv, P_slend_mean, P_Fk_1_prpn, P_slope, P_alt, card_4wayN, card_4wayNE, card_8way,MPI_100, MPI_200, MPI_500, MPI_1000, MPI_2000, WindSheltS1, WindSheltNE1, P_thinned, Estab_sph, Sph_drop, u_wind_tim, u_rain, u_min_temp, u_mint_rain, u_air_pr, u_rain_wind_tim, P_dbh_mean_NRML, P_tree_ht_mean_NRML, Tops_prpn_DAM, Plot_no, P_YOE, P_YOM, P_stand

model fitting	263 observations, 103 less than 1						
choice of variables	variables in model	variables at each split	R^2	mean	MAPE	bias: intercept	bias: slope
top ten variables by explanatory power	P_slend_mean, Tops_prpn_DAM, WindSheltNE1, P_tree_ht_mean_NRML, P_YOE, P_YOM, P_thinned, P_sph_equiv, u_min_temp, P_alt	4	0.097	0.969	3.3	0.864	0.108

fit statistics best model	Moran's I observed	Moran's I expected	p-value	autocorrelated
	-0.0185	-0.0039	0.1516	no

model validation	54 observations, 23 less than 1				
fit statistics best model	R^2	mean	MAPE	bias: intercept	bias: slope
	0.025	0.965	3.8	0.925	0.041

Figure 3-25, below, shows the relative importance of variables for this model.

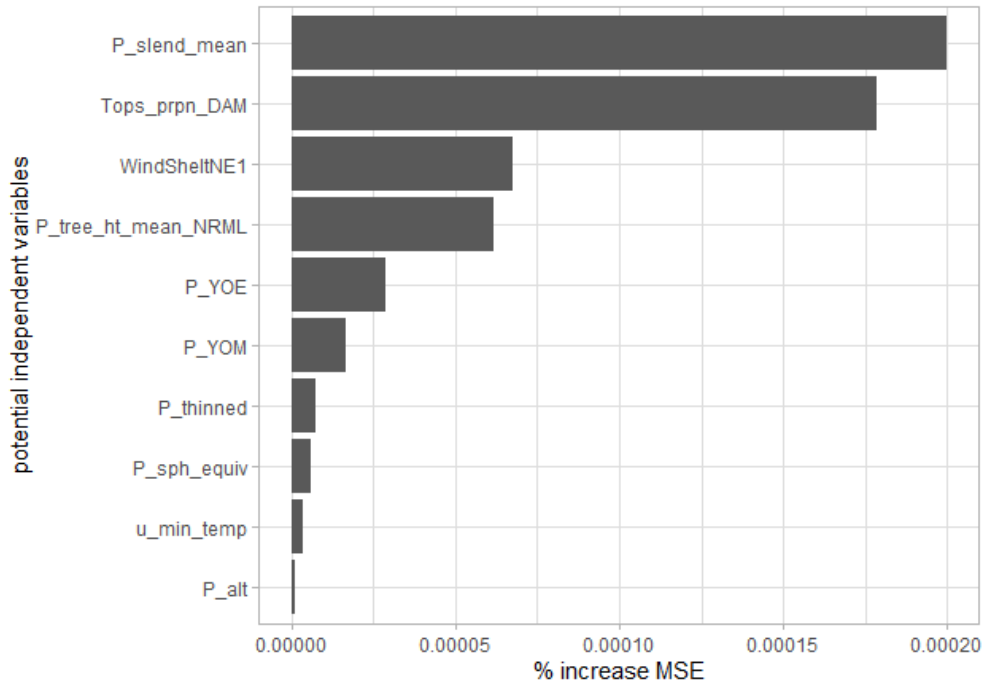

Figure 3-26: relative importance of variables for best random forest model for Douglas-fir Prpn_LIVE.

Plotting (Figure 3-27) the actual and predicted values for the fitting and validation steps data further underlines the wholly inadequate fit for this model.

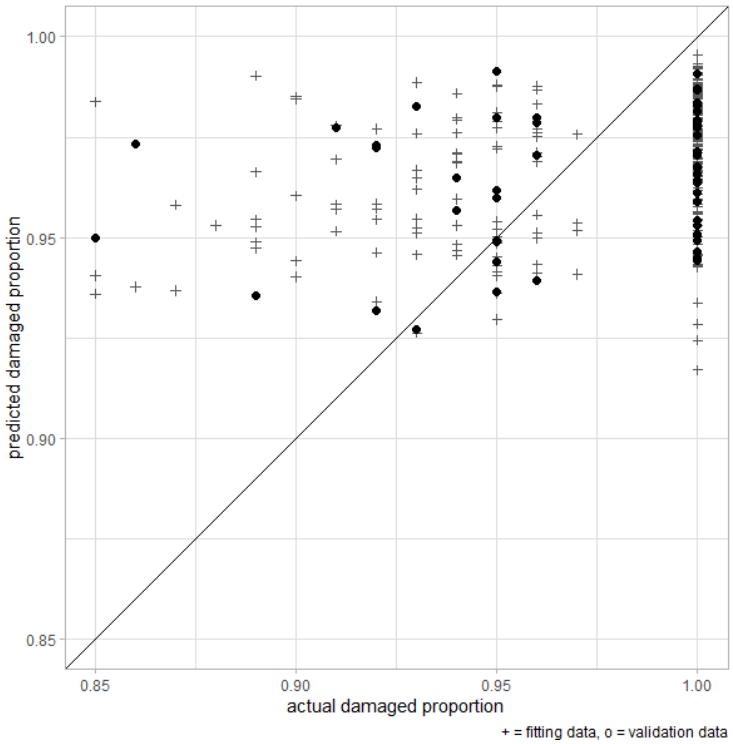

Figure 3-27: Visualising best random forest model for Douglas-fir Prpn_LIVE. Actual and predicted values from validation and fitting data; 1:1 line for reference.

4 Discussion

4.1 Summarised answers to the research questions

This section comprises summarised answers to the research questions. Detailed discussion of these results may be found in sections 4.2 to 4.6.

1. Do rates of tree damage differ significantly for radiata pine and Douglas-fir at Geraldine Forest?

Rates of tree damage do differ significantly for radiata pine (*Pinus radiata*) and Douglas-fir (*Pseudotsuga menziesii*) at Geraldine Forest. Douglas-fir has higher mean broken heights, lower proportion damaged, and a lower proportion of live trees than radiata pine (see section 3.1: *Establishing a difference between the species*). The lower proportion alive may be attributable to factors other than damage from wind and snow (see section 4.4.3), but the other two results probably result from a lower intrinsic vulnerability to damage for Douglas-fir.

2. How well can damage in radiata pine and Douglas-fir be modelled?

As presented in Section 3: *Results*, three of the models created have moderate explanatory power, defined as no significant autocorrelation in the residuals of the fitted model, fitting and validation R^2 >= 0.4, and slope of bias check for fitting and validation data >=0.35. Therefore, these three models warrant detailed discussion and comparison with previous literature. As listed in Table 3-2, these three models are:

- the linear regression model for radiata pine plot mean broken height (*P_tree_ht_mean_BRKN*)
- the random forest model for radiata pine plot mean broken height (*P_tree_ht_mean_BRKN*)
- the logistic regression model for radiata pine proportion of damaged trees per plot (*Tops_prpn_DAM*) for plots that have all tree tops assessed.

The lower explanatory power of the other models (see *Results* for these models, and also section 4.4) precludes the drawing of any conclusion other than the obvious: the explanatory variables do not explain much in these cases.

Table 4-1: List of models, showing those that receive further discussion and comparison.

model of	model by	sufficient explanatory power to warrant discussion and comparison?
***P_tree_ht_mean_BRKN*, radiata pine**	regression	yes
	random forest	yes
***P_tree_ht_mean_BRKN*, Douglas-fir**	regression	*no*
	random forest	*no*
***Tops_prpn_DAM*, radiata pine, all plots**	regression	*no*
	manual hurdle	*no*
	random forest	*no*
***Tops_prpn_DAM*, radiata pine, plots with all tops assessed**	regression	yes
	random forest	*no*
***Tops_prpn_DAM*, Douglas-fir**	regression	*no*
	random forest	*no*
***Prpn_LIVE*, radiata pine**	regression	*no*
	random forest	*no*
***Prpn_LIVE*, Douglas-fir**	regression	*no*
	random forest	*no*

3. Which modelling approach creates the most explanatory and least biased models?

On the whole, regression models had better explanatory power and less bias than random forest models, although the random forest model for plot mean broken height (*P_mean_ht_BRKN)* performed slightly better than the regression model for *P_mean_ht_BRKN.* Possible reasons for the worse performance of random forest models are discussed in section 4.7.1: *Comparison of regression and random forests*.

4. Which tree, stand and topographic conditions significantly affect tree damage?

Table 4-2, below, lists the explanatory variables for the three models with moderate explanatory power (variables not in any of the models are not shown). Topographic variables are useful in the regression models but not in the random forest model, and neither regression includes any weather variable. Definitions of the variables are given in section 2.4.2.

Table 4-2: variables found in three models with moderate explanatory power. (np) = not present in model, (na) = not applicable in model of this type.

	Radiata pine model for		
	P_mean_ht_BRKN by multivariate linear regression with mixed-effects	*P_mean_ht_BRKN* by random forest	*Tops_prpn_DAM, all tops assessed* by multivariate logistic regression with mixed-effects
plot description variables	P_stand (as a mixed effect) *(np)* *(np)* *(na)*	*(np)* P_YOM *(np)* *(np)*	P_stand (as a mixed effect) *(np)* *(np)* Plot_no (as a mixed effect)
tree description variables	*(np)* *(np)* P_age_meas_cs *(np)* *(np)* *(np)*	P_BA_ha_equiv *(np)* P_age_meas P_tree_ht_mean_NRML *(np)* *(np)*	*(np)* P_sph_equiv_cs *(np)* *(np)* P_dbh_mean_NRML_cs Prpn_LIVE_cs
silvicultural history variables	P_pru_prpn_cs (np)	P_pru_prpn P_pru_ht	*(np)* *(np)*
terrain variables	P_alt_cs card_4wayNE *(np)*	*(np)* *(np)* *(np)*	*(np)* card_4wayN MPI_200_cs
weather variables	*(np)* *(np)* *(np)* *(np)*	u_wind_tim u_air_pr u_rain u_rain_wind_tim	*(np)* *(np)* *(np)* *(np)*

5. Can the models developed be used to predict damage from new data?

The fourth major finding is that, due to the degree of bias, the imprecise predictions, and the reliance on mixed-effects models, the three models with moderate explanatory power should not be used to create numeric predictions of damage from new data.

6. Do the research findings suggest forest management practices that may reduce tree damage in radiata pine?

The fifth major finding is that results from the regression analyses do suggest some management measures to reduce tree damage by wind at Geraldine Forest. Major measures suggested are as follows (comprehensive discussion of these points is in section 4.6):

- consider planting Douglas-fir rather than radiata pine on the most damage-prone topography, which is north-east and south-east aspects, and/or situations with high values of the morphometric protection index.
- if planting radiata pine, choose a low final stocking: high stocking is associated with higher proportions of damage.
- if planting radiata pine, choose a short rotation: large tree diameters are associated with higher proportions of damage, reflecting both plot lifespan and time-independent effects of diameter.
- if planting radiata pine, consider pruning it, as pruned trees have higher broken heights (meaning more salvageable tree remains below the break), although this does run counter to the recommendation for a short rotation.
- if planting radiata pine, plant it at lower elevations, where growth is faster, allowing a desirable piece size in a shorter rotation.

4.2 Differing damage in radiata pine and Douglas-fir

Douglas-fir has higher mean broken heights, lower proportion damaged, and a lower proportion of live trees than radiata pine at Geraldine Forest (see section 3.1: *Establishing a difference between the species*), but is grown across a similar range of terrain (see Figure 2-1), and in a similar way (as single-species, single-age stands). Therefore, it appears that Douglas-fir is intrinsically less prone to wind and snow damage than radiata pine at Geraldine Forest. This point is reinforced when one considers that Douglas-fir is longer-lived than and was measured at a later age than radiata pine, and thus has more exposure to windy weather, but accumulates less damage (see plot age at measurement in Table 2-10 and Table 2-11.

This finding agrees with the findings of a study by Moore and Gardiner (2001), which tested the prevailing wisdom among forest managers that Douglas-fir is more wind-firm than radiata pine. Moore and Gardiner applied a mechanistic model to field and published data, and found that a baseline stand of Douglas-fir had a higher damage threshold windspeed (24.3 metres per second) than a baseline stand of radiata pine (20.6 metres per second). The probability of damage thresholds being exceeded during the stand's lifetime was 2.3 times greater for the radiata pine stand than for the Douglas-fir stand (0.115 versus 0.050), despite the much shorter modelled rotation of radiata pine (28 versus 45 years).

In the international literature regarding empirical models of wind damage to trees, species is often an important factor, but there appears to be no universal tendency regarding which species are most vulnerable, other than that deciduous angiosperms are less vulnerable to winter storms than conifers, due to their deciduous habit. For example (and assuming one can generalise responses to wind and snow to the genus level) Wright and Quine (1993) found that deciduous angiosperms were less damaged than *Picea* and *Larix*, which were less damaged than *Pinus.* However, Jalkanen and Mattila (2000) found *Betula* less damaged than *Pinus* and *Picea*, which had approximately the same levels of damage; and Veblen et al. (2001) found *Populus* to be less damaged than *Pinus*, which was less damaged than *Picea*, which was in turn less damaged than *Abies*. The only empirical study of wind damage frequency that included both *Pinus* and *Pseudotsuga* (Schmidt et al. (2010)) did not discern a significant effect of tree species.

Given the lack of informative content in the international literature, the agreement between this study's findings and that of Moore and Gardiner (2001), namely that Douglas-fir suffers less from wind damage than radiata pine, is valuable for the New Zealand context, particularly as the two studies have quite different modelling approaches but reach related conclusions. With these two findings, and supporting information, for example, the multiple anecdotal accounts in Somerville et al. (1989), it appears probable that Douglas-fir suffers less damage from wind and snow than radiata pine across all of New Zealand.

4.3 Discussion of models with moderate explanatory power

All three of the best-performing models show bias in both the model fitting and model validation steps, so that small values of the response variables are over-predicted and large values of the response variables are under-predicted; in other words, the models are under-fitted. It therefore seems likely that the models are to some degree mis-specified. In the regression models, care has been taken to check for variable significance (by p-values), variable meaningfulness and contribution to model power (by model coefficients) and variable independence (by variance inflation factors). Similarly, the random forest model was checked by ranking percentage mean square error by variable, to ensure that all included variables made a real contribution to reducing overall model error. Thus, it seems unlikely that the mis-specification is due to inclusion of irrelevant variables. It is more likely that that the models are mis-specified due to the lack of some influential variables. It is not possible to determine what these variables might be, apart from noting that they must be different to those variables trialled during model development (as listed in sections 2.6.6 to 2.6.8). The inclusion of mixed-effects in the regression models of radiata pine *P_tree_ht_mean_BRKN* and radiata pine *Tops_prpn_DAM* also points to the probable existence of influential variables that are missing from the explanatory variable set.

For the regression models, functional form mis-specifications are also possible. All variables were included as linear variables. The relationships between single explanatory variables and response variables in this research generally have correlations of less than 0.4 absolute, and visual checks of plots of these rather inexact relationships did not reveal any obvious candidate for modelling with, for example, transformed variables or variables or a power term. Interactions were also investigated during model development, but not were useful enough to warrant inclusion in final versions.

This study addressed spatial autocorrelation by trialling the inclusion of topographic variables in the models themselves, and addressed first-order conditional autoregressive structure (CAR1) trialling the use of the plot-level variables *P_stand*, *P_YOE* and *P_YOM* as mixed effects in regression analyses and as ordinary variables in random forests. This approach appears to have been successful, as regression models of radiata pine *P_tree_ht_mean_BRKN* and *Tops_prpn_DAM*, and the random forest models of radiata pine *P_tree_ht_mean_BRKN*, returned values for the Moran statistic (Moran's I) that indicate autocorrelation of the fitting dataset residuals is not significant.

Past studies generally use the presence or absence of damage to trees or plots, or the proportion of damage to plots or stands, which allows for comparisons between this research and past research for the response variable *Tops_prpn_DAM*. Very little of the literature examined for this study uses the *height* of breakage as a response variable for studying tree damage, so few comparisons are available for the response variable *P_mean_ht_BRKN*. Wrathall (1989) is the exception. This study, into the effects of a severe ex-cyclonic windstorm on 17 – 23 year old radiata pine trees, included modelling the actual broken height of the trees. The study also included the proportion of total tree height that was lost to breakage, through matching of broken-off tops to their original trees during field data collection.

An early idea for this research was to compare the variables correlated with wind and snow damage, as revealed by modelling, between the two species under study, with a view to investigating whether the variables were the same, similar, or different. This could have shed light on whether the correlates of damage are the same or different for radiata pine and Douglas-fir. However, none of the Douglas-fir models provided good explantory power, and so this comparison could not be made.

4.3.1 Model of plot mean broken height by linear regression for radiata pine

The results given in section 3.3.1.1 show that the linear regression model for *P_tree_ht_mean_BRKN* has a significant positive intercept of 14.5 m for the fitting dataset. This is the *expected* broken height for a plot on a north-east aspect when P_age_meas, P_pru_prpn, and P_alt are all at their (centred) mean values. The full-dataset mean of *actual* (from the field observations) broken heights is 14.5 m. The mean *predicted* broken height for model fitting is 14.5 m, and 14.4 m for model validation. That these figures are all so similar shows that the model predicts mean broken height very well, although the bias check figures show that predictions become less and less accurate further away from the mean.

According to the model coefficient for *P_age_meas_cs* (2.747), as the age of trees increases, the broken height also increases. Speculatively, trees may break at a weak point low in the live crown, possibly where a large branch or branches join the stem at a comparatively small stem diameter. As the age of tree increases, the (putative) weak points in the lower live crown will increase in height, leading to breakage which is at a reasonably consistent percentage of overall tree height, and hence the importance of tree age in this model. Tree age and mean unbroken tree height strongly correlate in this study's data (0.66), but tree age was the stronger predictor. As the total tree height at time of breakage is unknown, these contentions cannot be numerically supported; but see the photographs in the Appendices for examples of wind damage at Geraldine Forest, and also the findings of other research (in the final paragraph of this section) which support these ideas.

As the proportion of trees pruned (*P_pru_prpn_c*) in a plot increases, so too does the height at which trees break (the model coefficient is 1.265). This relationship has a relatively high standard error in comparison to the coefficient, probably because the distribution of *P_pru_prpn_c* has modes at both 0 (no pruning) and 1 (all trees pruned). Because of this bimodal distribution (shown in Figure 6-43), consideration was given to including pruning in the model via *P_pruned*, a binary categorical variable, but *P_pru_prpn* gave better model performance. It is possible that pruning offers some degree of protection against breakage, by reducing the crown area on which wind can act, or by removing large branches against which snow accumulates to breakage-causing weights. The empirical studies of wind and snow damage cited in Table 1-1 do not include pruning as a factor, but anecdotal accounts of the protective effect of pruning occur in the New Zealand literature. For example, Turner (1989) observed that in Lake Taupo Forest, New Zealand, pruned stems survived an ex-cyclone in 1988 more frequently than unpruned stems, and speculated that this was due to pruned stems having reduced sail area, and/or because the most vigorous trees had been selected for pruning. Likewise Olsen (1989) observed that in young stands (aged 5 - 9 years) that had been pruned but not yet thinned, damaging winds would sometimes blow over the unpruned stems, but leave the pruned stems standing.

Aspect (*card_4wayNE*) has some influence in this model, and is represented here by the four cardinal directions of north-west (coefficient 0.832), north-east (no coefficient as it is the base level for the category), south-east (coefficient -0.526), and south-west (coefficient -0.529). Significance checks established that north-west aspects have the highest broken heights, with south-east and south-west not being significantly different from the base level of north-east, or from each other. Interpretation of the relatively small effects of *card_4wayNE* is difficult without further analysis and is highly speculative. The situation as modelled may be consistent with confounding between aspect, tree height, and lee slopes, with larger trees occurring on north slopes and smaller occur trees on south slopes, due to differences in radiation received and air temperature, and lee slopes suffering stronger wind effects. The existence of lee slope effects on east-facing slopes at Geraldine Forest is suggested by the discussion of past wind records (section 4.7.2), and the fact that *card_4wayNE* also occurs as a

significant factor in the model for the proportion of damaged tops per plot (see *Results* section 3.4.1.3 and *Discussion* section 4.3.3). No analysis has been undertaken to establish that trees are in fact larger on north slopes.

The coefficient (-0.954) for *P_alt* indicates that as the elevation of a plot increases, the height at which trees break decreases. If trees frequently break at some consistent point of canopy architecture, as contended in the discussion of age in this model, then breakage may be lower at higher elevations because the trees are smaller at higher elevations, giving a lower absolute height of the consistent point. In other words, there may be a confounding of the effects of age and elevation. To examine whether trees are in fact smaller at higher elevations, some third-party data were available[3]. These data are based on the same LiDAR imputation plots and the same secondary (15 m resolution) digital elevation model as used in this study, and include the site index for those plots. The site index for radiata pine is defined as the mean top height of trees at age 20 years (Goulding, 2005). Data at a range of ages can be the basis for calculating site index, meaning that site index has the useful property of being age-invariant, and thus suitable for illustrating the effect of topographic influences on tree height. Figure 4-1, below, clearly shows the negative correlation between elevation and site index at Geraldine Forest, supporting the contention that trees are smaller at higher elevations.

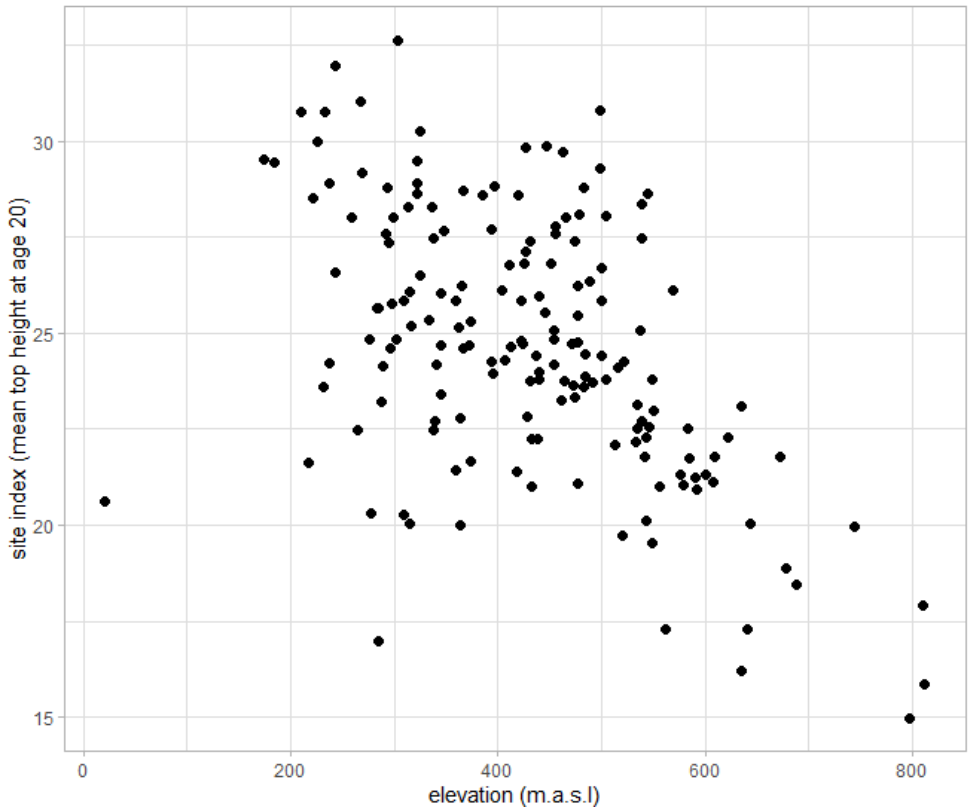

Figure 4-1: site index by elevation, Geraldine Forest

When considering the data hierarchy in this study, it is apparent that data from plots in a single stand are very likely to have first order conditional autoregressive structure (CAR1), because stands (by

[3] used with the permission of Prof. Euan Mason, the author of the dataset

definition) are planted in a bounded geographic area, have a common establishment date and have the same silvicultural history. Using *P_stand* as a mixed-effect in this model was in important in compensating for CAR1: all model versions trialled without *P_stand* as a mixed effect it returned a significant value for the Moran statistic. In addition, using *P_stand* in this model gave a noticeable increase in model explanatory power (see the marginal and conditional R^2 results for this model in section 3.3.1.1).

Comparing this result to the literature introduced in Chapter One is challenging, because very few of those studies examine the heights at which trees break. However, the mean predictions of 14.5 m (fitting) and 14.4 m (validation) may be compared with Knowles and Paton (1989), who found the majority of broken heights at the Tikitere Agroforestry trial fell between 10 and 14 m; those trees were 15 years old, versus a mean age of measurement in this study of 24 years. That study also found that trees broke at about 60% of their original height and always at a branch whorl; there was considerable variation in the location of breakage, but breakage in the top two metres of stem or at the top of the pruned stem was very uncommon. Wrathall (1989) found that trees broke at about 38% of their original height, again with considerable variation, and that taller trees break at higher heights, both in absolute terms and as a proportion of original height. Absolute and proportional heights of breakage had positive correlations with dbh^2, original tree height, and crown size (crown depth multiplied by crown height). In this study, although the mean diameter of normal trees (*P_dbh_mean_NRML*) has a noticeable positive correlation of 0.24 with broken height, diameter was not a significant explanatory variable when trialled and so did not warrant inclusion in the final model discussed here. The latter two variables were available because broken-off tops were matched to their original trees during the study fieldwork, a process not undertaken for this study.

4.3.2 Model of plot mean broken height by random forest for radiata pine

Because random forests do not have model coefficients to examine, when discussing the relative contribution of a model's explanatory variables, it is helpful to present the correlations between the response and explanatory variables. Table 4-3, below, shows these correlations for *P_tree_ht_mean_BRKN* and the explanatory variables. It is also helpful to plot the ranked importance of the variables, as in Figure 4-2, which shows the expected increase in mean square error of prediction if the variable is excluded from the set available to the *cForest* algorithm.

Table 4-3: correlations among variables for P_tree_ht_mean_BRKN by random forest. Calculated from the model fitting data. The categorical variable P_YOM cannot be included.

	P_tree_ht_mean_BRKN (response)	P_age_meas	P_pru_prpn	u_wind_tim	P_tree_ht_mean_NRML	P_BA_ha_equiv	P_pru_ht	u_rain	u_air_pr	u_rain_wind_tim
P_tree_ht_mean_BRKN (response)	1.00									
P_age_meas	0.29	1.00								
P_pru_prpn	0.21	0.00	1.00							
u_wind_tim	0.28	0.95	-0.02	1.00						
P_tree_ht_mean_NRML	0.26	0.45	0.12	0.43	1.00					
P_BA_ha_equiv	0.24	0.28	0.02	0.27	0.26	1.00				
P_pru_ht	0.07	-0.20	0.27	-0.20	-0.01	-0.10	1.00			
u_rain	0.28	0.95	0.01	0.90	0.45	0.28	-0.22	1.00		
u_air_pr	0.28	0.93	0.02	0.91	0.46	0.28	-0.24	0.91	1.00	
u_rain_wind_tim	0.22	0.68	0.02	0.65	0.46	0.29	-0.28	0.72	0.74	1.00

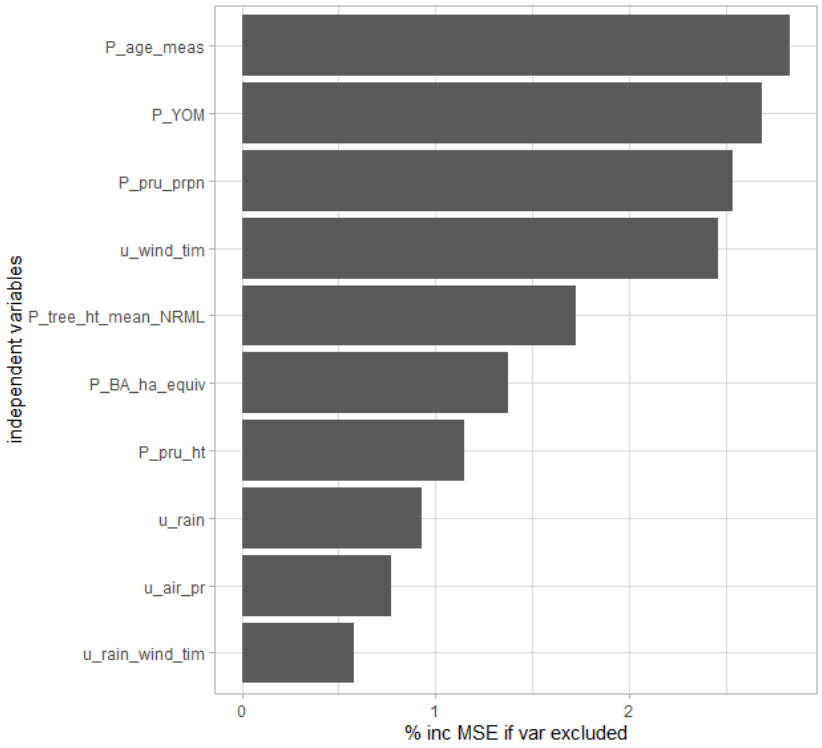

Figure 4-2: relative contributions of variables included in best random forest model for radiata pine P_tree_ht_mean_BRKN

The results given in section 3.3.2.1 show that the random forest has a mean of 14.5 m for *predicted* broken heights from model fitting, and a mean of 14.6 m from model validation. The *actual* mean broken height across all plots is 14.5 m (as noted in section 4.3.1). These values are very similar, and

the same points apply as for the discussion of means in section 4.3.1 – including the point that the mean broken height alone is not especially useful. Although the model results show an overall bias (see Table 3-5 for an assessment of bias in this model), it is worthwhile to consider the contribution of each of the variables in the model.

P_age_meas is one variable that this model has in common with the regression model for radiata *P_tree_ht_mean_BRKN*. It is also the most influential predictor in both model this model and the regression model for radiata *P_tree_ht_mean_BRKN*. It is probable that similar reasons for its inclusion here apply as discussed in section 4.3.1, namely weak points associated with large branches low in the live crown, the height of which increases as the tree age. Given that this variable appears in two models that have entirely different theoretical precepts, it is highly likely that *P_age_meas* has a real and strong influence on predictions of *P_tree_ht_mean_BRKN*.

P_YOM, the plot measurement year, is the second most influential variable. This suggests that there is some variation in measurement technique between years, despite the standardisation efforts made during data compilation for this study. *P_YOM* does not appear in the comparable regression model (section 4.3.1): it was promising as a mixed-effect during model development, but was ultimately outperformed by *P_stand*. *P_stand* has similarities to *P_YOM*, since in this study the all the plots pertaining to one stand are created and measured together.

The third most influential variable in this random forest model is *P_pru_prpn*, the proportion of trees pruned per plot, which is another variable in common with the regression model for radiata *P_tree_ht_mean_BRKN*. For the same reasons as discussed in section 4.3.1 above, this model result may be expressing that pruning offers some degree of protection against damage; and again, as the variable appears in both model types, it is very likely to be a real and strong effect.

The fourth most influential variable is *u_wind_tim*, and alongside *u_wind_tim* one should consider *u_rain*, *u_air_pr*, and *u_rain_wind_tim*, which are 8th, 9th, and 10th most influential, respectively. Figure 4-2 and Table 4-3 show that a large proportion of the model power comes from this group of variables, which are highly with each other and with *P_age_meas*: by comparison, no weather variables are in the regression model for *P_tree_ht_mean_BRKN*. The inclusion of *u_wind_tim, u_rain* and *u_rain_wind_tim* may mean these variables are predictive due to the influences (high winds, wet soils, the combination of high winds and wet soils) proposed for these variables in *Methods*. However, the inclusion of this group of variables may alternatively mean that there remains some degree of over-reliance of correlated variables, despite the *cForests* algorithm having been designed to address this issue. The inclusion of *u_air_pr*, intended to be a proxy for days of poor weather *without* a link to a specific effect of that weather, is suggestive of some such issue. This author suggests exercising caution during any similar research into damage to trees that uses any random forest algorithm and returns a group of similarly highly correlated variables.

P_tree_ht_mean_NRML, the height of trees with normal tops, is the fifth most influential variable in this model. The correlation of *P_tree_ht_mean_NRML* and *P_age_meas* is 0.45 (Table 4-4, above), indicating that these variables are moderately related, but not the same. *P_tree_ht_mean_NRML* might express the effects of elevation on growth: as noted in section 3.4.1.3, trees are shorter at higher elevations. Alternatively, this could be another instance of the over-use of correlated variables; it is difficult to say which is more likely.

The sixth most influential variable in this model is basal area, *P_BA_ha_equiv*, which has a correlation of 0.24 with the response variable *P_tree_ht_mean_BRKN.* Basal area adds together the cross-sectional area of trees at breast height, to give a measure of site occupancy, which increases over time

unless some event removes stems, such as (planned) thinning or (unplanned) windthrow. In this research, basal area has been made comparable across plots by division by the plot area, which yields the per-hectare equivalent basal area. Assuming that the effect of basal area is not a re-expression of increasing tree age (their correlation is weak at 0.28 as shown in Table 4-3 above), then perhaps the inclusion of basal area expresses a mutual sheltering effect among trees on more highly occupied sites. By way of comparison, note that the regression model for radiata pine *P_tree_ht_mean_BRKN* (discussed in section 4.3.1) includes increases in mean broken height with increases in both of mean diameter (of unbroken trees) per plot and plot stocking. As *P_BA_ha_equiv* is ultimately calculated from the diameter of and number of stems, this constitutes a similarity between the models – and perhaps some assurance that *P_BA_ha_equiv* is a meaningful inclusion in this model.

Pruned height, the seventh most influential variable in this model, is predictive in addition to the effect of including *P_pru_prpn*. Figure 4-3 below shows that various plot mean pruned heights exist. It is possible that any protective effects of pruning against damage are less for lower pruned heights.

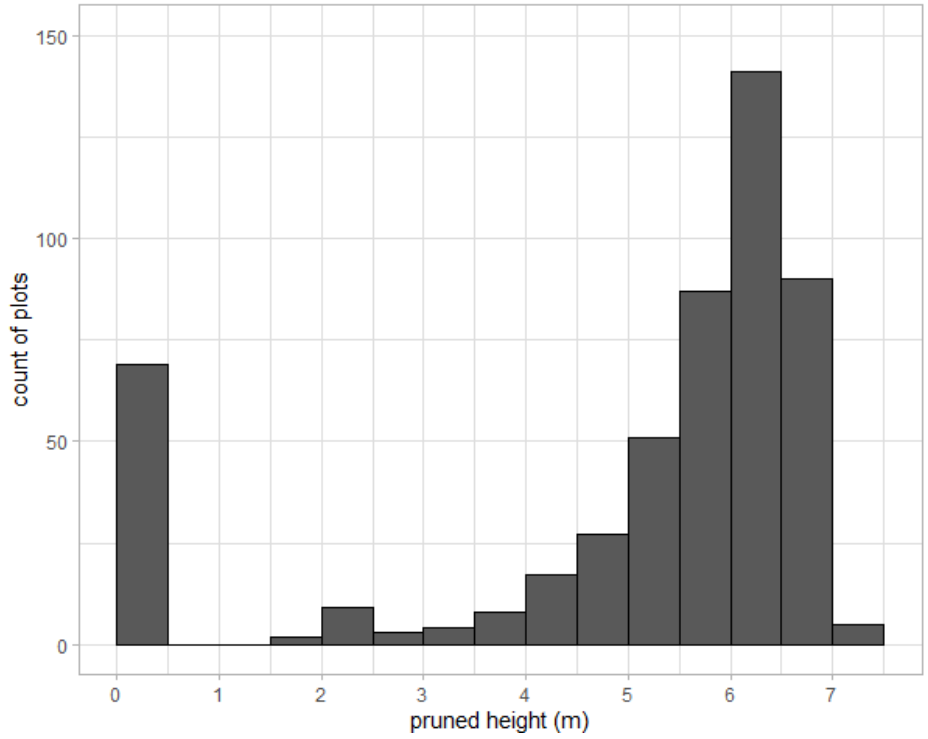

Figure 4-3: distribution of pruned heights in the radiata pine fitting dataset, mean: 5.0m, median: 6.0 m (the 0 to 0.5 m bar represents unpruned trees).

Comparing the model fit statistics for the regression and random forest models of *P_tree_ht_mean_BRKN* (Table 4-4, below) shows that the models perform similarly. If the inclusion of correlated predictors, as shown in Table 4-3, was adversely affecting the random forest, a poor performance in comparison to the regression model should occur; but no such difference in performance exists. Therefore, the *cForests* algorithm may have dealt with these correlations appropriately.

Table 4-4: comparison of model fit statistics for radiata P_tree_ht_mean_BRKN as modelled by regression and random forest

model type	data	R^2	MAPE	bias: intercept	bias: slope
regression	fitting	0.372 (fixed effects) 0.426 (full model)	24.5	8.919	0.386
	validation	0.414	23.7	8.714	0.407
random forest	fitting	0.422	25.2	8.902	0.392
	validation	0.527	20.2	7.949	0.477

4.3.3 Model of proportion of damaged trees per plot with logistic regression for radiata pine, only plots with full tops assessment

Despite the fewer data points, which occur because only a subset of plots received a full tops assessment, this model for the proportion of damaged trees per plot (*Tops_prpn_DAM*) has improved explanatory power compared to the model for *Tops_prpn_DAM* model with all plots included. To compare, see Table 3-2*: Comparison of model performance on fitting and validation data.*

There are two potential reasons for this improvement. First, there is no possibility of incorrect values of the response variable proportion of damaged trees per plot (*Tops_prpn_DAM)*. Second, the mixed-effects logistic regressions performed in this research appear to have difficulty accommodating true zeros in response variables, but the subset of plots that had all tops assessed includes only four cases out of 198 with zeros for *Tops_prpn_DAM*.

The intercept of this of this model (-0.393) represents the mean proportion damaged when all the other numeric predictor variables are held at their centred means, i.e. zero. While Table 3-9 gives the raw model coefficients, and the coefficients expressed as percentage changes in the odds of damage, for the model intercept it is more informative to calculate the expected damage as a proportion: this is analogous to the examination of the mean undertaken in the first paragraph of section 4.3.1.

Transforming the model intercept of -0.393 to a proportion gives the figure *1/(1+1/exp(-0.393))***,** or 0.403. This is the *expected* proportion damaged for a plot with north-east aspect, when stocking (*P_sph_equiv_cs*), morphometric protection index at a 200 m horizon (*MPI_200_cs*), plot mean diameter of normal trees (*P_dbh_mean_NRML_cs*) and the proportion of live trees (*Prpn_LIVE_cs*) held at their centred means (zero). The full-dataset mean of *actual* (from the field observations) damaged proportions in radiata pine is 0.432. The mean *predicted* damaged proportion for model fitting is 0.387, and 0.372 for model validation. In a perfectly performing model, all these figures would be the same. That they are not shows points to some bias and overall inaccuracy in this model's results.

Turning now to an examination of the individual variables comprising the models, one finds that the plot per-hectare equivalent stocking, *P_sph_equiv_cs*, has first-equal influence in this model. The odds of damage increase by 35% for every standard deviation of increase in stocking. A possible interpretation is that trees in stands at higher stockings are relatively more slender for their height, and therefore more prone to breaking. The higher stocking may also increase the chances that trees break or fall from being struck by other broken or fallen trees, as the trees at higher stocking are physically closer together. Finally, a higher stocking could mean a reduced selection ratio at the time of thinning, with consequently fewer opportunities to remove stems that are already damaged at the time of thinning. By comparison, the data available for current and historic permanent sample plots

revealed that 14 permanent sample plots had dead, broken or defective tops (PSP code DT) at measurements that occurred before thinning, indicating tree damage before thinning is possible at Geraldine Forest.

This finding of increased damage with increased stocking agrees with other studies that utilised logistic regression to investigate forest damage. For example, Jalkanen and Mattila (2000) found high stocking was positively related to the incidence of wind damage in a nation-wide study in Finland; Mitchell et al. (2001) found similarly for severity of winter storm damage to managed stands on Vancouver Island, Canada; and Bennett (2002) found high stocking was positively related to the proportion of broken tops in radiata pine in Otago, New Zealand. However, it disagrees with the findings by Knowles and Paton (1989), who found that data from the Tikitere Agroforestry trial suggested no relationship between damage proportion and stocking in sheltered situations, and a *decreased* incidence of damage with increased stocking in exposed situations.

MPI_200_cs, the morphometric protection index at a 200 m calculation horizon, is also highly influential in this model: the odds of breakage increase by 35% for every standard deviation of increase in morphometric protection index. As higher values of MPI indicate lower values of topographic shelter, a reasonable assumption is that the morphometric protection index expresses the higher exposure to wind experienced by trees that are less sheltered. This is similar to studies using logistic regression by Mitchell et al. (2001) and Hanewinkel et al. (2014), who found a link between damage and high topographic exposure; and Lindemann and Baker (2002) and Lanquaye-Opoku and Mitchell (2005), who found a link between damage and high wind exposure. Martín-Alcón et al. (2010), when using linear regression with an auto-covariate, discovered that high topographic exposure is predictive of high damage.

Aspect has some influence in this model. As a categorical variable, *card_4wayNE* represents a change in the model intercept for each level of the variable. The cardinal directions north-west and south-west are significantly different to one another, and both are significantly different to the directions north-east and south-east, which are not significantly different to one another. Plots with north-west and south-west aspects have 18% and 27% decreases in the odds of damage, respectively. Some previous studies have found an influence of aspect when studying wind damage, including Lindemann and Baker (2002) and Lanquaye-Opoku and Mitchell (2005) by logistic regression, Schmidt et al. (2010) by generalised linear models, and Dobbertin (2002) by cross-validated classification trees. Given that the strongest winds the Geraldine Forest area (see section 4.7.2) occur from directions between south-east and north (when considered clockwise), it seems probable that this represents a lee slope effect of increased damage, relative to windward slopes. This matches the findings of Martin and Ogden (2006), in their review paper of wind damage to trees in New Zealand, and may be due to the increase turbulence on leeward slopes, as posited by those authors. However, the effects of lee slopes on wind damage to trees are by no means agreed upon in the international literature: see the discussion in Everham and Nicholas (1996).

This model includes *P_dbh_mean_NRML_cs*, the mean diameter at breast height of normal trees, with a 26% increase in the odds of breakage for every standard deviation of increase in *P_dbh_mean_NRML_cs*. Diameter is not especially common as a logistic regression predictor of snow and wind damage in the literature surveyed, but Jalkanen and Mattila (2000) discovered a relationship between increasing diameter (and also increasing stand age) and increasing damage, among data from a single-period forest inventory in Finnish forests exposed to both wind and snow. Díaz-Yáñez et al. (2017) discovered a relationship between increasing damage and increasing diameter (and also increasing tree height) with data from a forest inventory with four return visits, in Norwegian forests in affected by wind and snow.

It seems likely that *P_dbh_mean_NRML_cs* is expressing the lifespan of a plot and hence its accumulated opportunities to experience damage-causing winds and/or snow as the years pass. This may not be the sole effect of *P_dbh_mean_NRML_cs*, however: if it only expresses plot lifespan, then *P_age_meas* should have been equally useful as an alternative variable in the model, but it was not. The two studies cited above, which both find a tree lifespan variable (stand age and tree height respectively) influential in addition to diameter reinforce this idea.

The final variable included in the fixed-effects portion of this model is *Prpn_LIVE_cs*, the proportion of live trees per plot. As the proportion of live trees in a plot increases by one standard deviation, the odds of damage decrease by 23%. From this, one might conclude that some trees that fall or break go on to die. Live proportion as a predictor of proportion damaged was not found in the literature using logistic regression to predict wind and snow damage; Veblen et al. (2001) use tree live/dead status in chi-square and ANOVA analyses of damage in classified percentage groups (>75%, 50-75%, 3 patches 25-50%). The correlation between *Prpn_LIVE* and *Tops_prpn_DAM*, and the reasons for not creating a composite variable expressing both, is discussed in section 4.7.4.

As was the case for the model of plot mean broken height (*P_tree_ht_mean_BRKN*: section 4.3.1), using *P_stand* as a mixed-effect was vital in controlling for first-order conditional autoregressive structure: all model versions trialled without *P_stand* returned a significant value for the Moran statistic (see results for this model in section 3.4.1.3).

The inclusion of the observation-level random effect (OLRE) *Plot_no* was useful in this regression, giving a substantial contribution to explaining variance in this model (again, see section 3.4.1.3). This indicates the presence of over-dispersion (unexplained variance in the residuals) for this model: the over-dispersion statistic for a model with the same fixed effects as this model but no mixed effects is 1.96 (see Appendix 6.6 for details of that model). The presence of over-dispersion indicates that there probably are influential variables for this model that are not included in the explanatory variable set.

4.4 Factors reducing model explanatory power

4.4.1 Explanatory variable applicability and dataset size

In this study, the Results show that models of plot mean broken height (*P_tree_ht_mean_BRKN*) have higher explantory power than models of proportion damaged per plot (*Tops_prpn_DAM*), which in turn have higher explantory power than models of proportion of live trees per plot (*Prpn_LIVE*). This suggests that the explanatory variables studied here have most relevance to plot mean broken height and least to live proportion per plot. There also appears to be an effect due to the size of the model-building dataset: Douglas-fir models, which have a dataset approximately half the size of the radiata pine dataset (317 versus 625 plots), always exhibit less explanatory power than radiata pine models.

4.4.2 Inaccuracy in the response variable proportion of damaged trees per plot

An important distinction is between models that a) use data from all plots, and that b) use only a subset of data from plots in which all the trees received an assessment of their tops. Radiata pine has such a subset; Douglas-fir does not. When comparing model explanatory power, models of proportion damaged per plot (*Tops_prpn_DAM*) of type a) performed worse than models of type b). This occurs for both logistic regression and random forests. This suggests that the technique for estimating whole-proportion damaged per plot from a sample of tops assessed in each plot is to some degree inaccurate, a possibility discussed in detail in section 2.6.4.1.

4.4.3 Model formulation issues

Imbalance in the explanatory variables, both in numeric and geographic terms, may be affecting some models. The strongest example is how *P_thinned* dominates the regression of *P_tree_ht_mean_BRKN* for Douglas-fir, so that the model responses occur in two bands, which are clearly visible in Figure 3-2. All the unthinned plots are from one stand. As stands comprise a specific geographic area, planted in the same year and receiving the same silviculture, the explanatory variable data for the unthinned plots in this regression have very narrow ranges, rendering the model suspect.

Both logistic regression and random forest models do not predict zeros accurately in models of proportion damaged per plot (*Tops_prpn_DAM*) and proportion of live trees per plot (*Prpn_LIVE*). *Tops_prpn_DAM* includes zeros and the *Prpn_LIVE* includes ones, which are the inverse of zero proportion dead. For these variables, no model predicts zeros when the true value is zero, nor ones when the true value is one; although the difference is small in some cases, in others it is substantial. Further complicating the interpretation of zeros is the clustering at zero in the response variable *Tops_prpn_DAM* when all plots are included, as discussed in section 4.4.2, above. This undesirable phenomenon lay behind the attempted use of hurdle models and zero-adjusted binomial models. Poor predictions of zero also occur in the literature examined. For example, in a study in Sweden, which attempted to predict the presence or absence of wind and snow damage in permanent sample plots in *Pinus sylvestris,* (Fridman et al., 1998) found that logistic regression analyses predicted a fewer plots to have zero damage than the actual occurrence of zero damage.

Although the attempted of hurdle models and zero-adjusted binomial models did not give large advances in model explanatory power, the probabilistic-then-predictive pattern of these models did

clarify that the examination of damage to trees at Geraldine Forest is probably affected by the imbalanced classification problem. For the response variable *Tops_prpn_DAM,* there are two classes of trees at Geraldine Forest: those that break (or blow over), and those that do not. For the response variable *Prpn_LIVE,* the two classes are that trees die, and trees that do not. The outcome for a plot of trees is an accumulation of these binary events. However, the summary statistics shown in Table 2-10 (radiata pine) and Table 2-11 (Douglas-fir), and also model outcomes for the binary step of the manual hurdle model (Table 3-8) make it clear that that data are very imbalanced. Cases of the problem being modelled, namely tree damage or death, are rare in relation to the cases where this does not happen: the classes are imbalanced. It can be difficult to predict outcomes for members of the rarer class, and this problem has given rise to many discussions and techniques: see for example Ali, Shamsuddin, and Ralescu (2015), or Maalouf and Siddiqi (2014). In this study, modelling was at the plot level, partly because many of the potential explanatory variables are not meaningful at the tree level: for example, stocking only exists for groups of trees. In addition, the exact location of trees in plots is unknown, so there are no tree-specific values for topographic variables. While predicting at the plot level allowed the inclusion of a wider range of explanatory variables, it is probably not a coincidence that the least satisfactory models in this study were for *Prpn_LIVE*, where the class 'tree is dead' is very rare.

Another possible reason for the very poor predictive ability of models of *Prpn_LIVE* is a partial violation of the assumption mentioned in the Introduction, namely that the damage metric, tree death in this case, is actually influenced by wind or wind plus snow damage. The correlations between *Tops_prpn_DAM* and *Prpn_LIVE* are discernible but not especially strong (see 3.5), which suggests that another cause of tree death is present. This could well be intraspecific competition among the trees for growth resources, such as light and water.

Model of *Prpn_LIVE* would more usually be called mortality models in New Zealand forestry parlance, and mortality models for radiata pine usually take quite a different format, which is purely empirical, not attempting at all to address the causes of tree death. As described by Woollons (1998), the most common and successful sort of mortality predictions in even-aged radiata pine stands are created by fitting difference equations, which calculate the stocking (live trees) and mortality (dead trees) at time 2 from the stocking at time 1 and the period elapsed between time 2 and time 1. The equations are fitted to data from plots of trees that are measured repeatedly to yield time-series data. The models created here for *Prpn_LIVE* are of a quite different format, as the data for this study are not time-series data, and are evidently nowhere near as suitable as the more conventional approach. Using the proportion of live trees as a response variable appears uncommon in the literature, although Everham and Nicholas (1996) in a review article suggest that compositional loss (percentage stems dead), along with structural loss (percentage stems damaged) should always be included as one of the measurements collected in surveys after wind damage.

4.5 Predictions of damage in new areas or from new data

4.5.1 Applicability of created models to new or future data

To discuss the applicability of the models created during this study for new data, for example, data from future stand inventories, it is necessary to consider two points: the bias in model predictions, which applies to all models, and the nature of mixed-effects regression models.

All models created in this study exhibit bias. Predictions of low values of response variables are too high, and predictions of high values of response variables are too low. Re-use of the models with new or future data will therefore also contain bias. This applies even to the three models with moderate explanatory power.

In this study, the plot description variables *P_YOE*, *P_YOM* and *P_stand* were trialled during model creation as random-intercept mixed-effects, along with the *Plot_no* as an observation-level random effect (OLRE), as described in section 2.6.2. Mixed-effect models can predict new data, but if the levels of the random effects groups in the new data were not present at model building, the prediction for those new levels will be for an 'average' group, where 'average' is the mean of the distribution of the group. One may observe in the Results that all the logistic regression models, which all use *Plot_no* as an OLRE, have a drop in explanatory power between the fitting and validation datasets, precisely because the validation datasets constitute new values of *Plot_no*.

This means that if the model for radiata pine *Tops_prpn_DAM* is applied to new or future data with new values of *Plot_no* (inevitable for new data) or *P_stand* (possible for new data), the contribution from the mixed-effects *Plot_no* and *P_stand* will be for average groups. Therefore, the predictions will contain mostly information from the fixed portions of the model. Similarly, if the model for radiata pine *P_tree_ht_mean_BRKN* is applied to data with new values of *P_stand* (possible for new data), the contribution from the mixed-effect *P_stand* will be for the average group, and the predictions will contain mostly information from the fixed portions of the model.

These points taken together mean that the models presented in this study are not directly useful for producing numeric predictions of future damage.

4.5.2 Forest-wide spatially explicit predictions of damage

An early concept for this study was to create models of plot mean broken height (*P_tree_ht_mean_BRKN*), proportion damaged per plot (*Tops_prpn_DAM*) and proportion of live trees per plot (*Prpn_LIVE*), and then deploy these across the entire Geraldine Forest landscape in a manner independent of existing tree crops, to create a lasting and spatially explicit raster-based reference for the likely degree of damage.

This was not possible, firstly because even the three most satisfactory models created include bias and mixed effects rendering them unsuitable for re-use. The second reason is less immediately obvious, but equally important. Those same three models all include variables that only arise from measurements of trees. Experimentation during model formulation showed that no satisfactory models could be created that contained only the variables that a forest manager could know or reasonably assume in the absence of measurements of trees, which are a) the topographic variables, and b) the tree description variables *P_age_meas*, *P_sph_equiv*, *P_thinned*, *P_pruned*, and *P_pru_ht*. Consequently, full-wide spatially explicit estimations of damage are not possible.

4.6 Management recommendations

All these management recommendations assume that the past is a reasonable guide to the future: the factors influencing damage to trees, as discovered during modelling of past data, are taken to be applicable to future data. This assumption could be wrong; but there are no data to assess this.

This study has established that damage to trees does differ significantly between radiata pine and Douglas-fir at Geraldine Forest. As presented in section 3.1, Douglas-fir has higher mean broken heights, lower proportion damaged, and a lower proportion of live trees than radiata pine. The lower proportion damaged occurs despite the Douglas-fir plots being much older on average (40.0 years versus 23.9 years) than the radiata pine plots, and thus having been exposed to more potentially damaging wind and snow events. If growth of an intact crop (or nearly intact crop) is important, Douglas-fir is the superior choice. From a management perspective, however, this causes a trade-off with the much longer rotation of Douglas-fir.

Results regarding the proportion of damage per plot (see sections 3.4.1.3 and 4.3.3) establish that that the most damage-prone topographic situations for radiata pine are north-east and south-east aspects, and/or situations with high values of the morphometric protection index, which equates to low topographic shelter. Forest managers could choose to plant Douglas-fir instead of radiata pine in these areas.

If radiata pine is planted, it may be desirable to plant it at lower elevations. Because growth is faster at lower elevations (as discussed in section 4.3.1), it may be possible to obtain a desirable piece size in the shorter rotation suggested above.

For radiata pine, managers could choose a low final stocking, because higher stocking is associated with higher damage (see section 3.4.1.3 and section 4.3.3). A short rotation may also be beneficial, as larger diameter, here acting as a metric of tree lifespan, is associated with higher damage. Also, larger trees are more prone to breakage, additional to the effects of lifespan.

However, the recommendations for low final stockings should be taken in the context of the prevailing wisdom about what silviculture to apply to radiata pine to mitigate wind damage; for example, Somerville (1995) in his review article 'wind damage to New Zealand state plantation forests', gives heavy and late thinning as increasing the risk of wind damage. Ledgard (1982) observed increased damage after thinning for Geraldine Forest, both anecdotally and from studying the progress of a trial of different thinning and pruning regimes, and further suggested that regimes involving two thinnings showed increased damage levels. Therefore, while low final stockings are recommended, thinning used to achieve this should not be especially heavy or late, and preferably as a single operation.

In general, only the part of a tree below a break will be harvested; therefore, the height at which trees break is of interest. The results suggest that radiata pine will break at taller heights if pruned, so consideration should be given to pruning radiata pine. Admittedly, there will be a conflict between the recommendation for short rotations and the recommendation for pruning.

4.7 Other findings

4.7.1 Comparison of regression and random forests

Overall, regression models performed better than random forest models, returning higher R^2, and lower MAPE and bias values. Table 4-5, below, summaries this situation (to see the individual performance metrics, refer to results Table 3-2). The lower R^2 and higher bias figures lead to the conclusion that the random forest models are under-fitted; they do not explain the variation in the data. The regression models are also under-fitted, but to a lesser degree.

Note that 'better model performance' is only a comparison between the model types for each variable, not an indicator of model explanatory power. As discussed in section 2, the only models with moderate explanatory power are the linear regression model for radiata pine *P_tree_ht_mean_BRKN*, the random forest model for radiata pine *P_tree_ht_mean_BRKN*, and the logistic regression model for radiata pine *Tops_prpn_DAM* for plots with all tops assessed. All other models have low or very low explanatory power.

Table 4-5: best-performing model type by species and response variable.

variable	**species**	**data**	**better model performance from**
P_tree_ht_mean_BRKN	radiata pine	fitting	random forest
		validation	random forest
	Douglas-fir	fitting	regression
		validation	regression
Tops_prpn_DAM	radiata pine (all plots)	fitting	regression
		validation	random forest
	radiata pine (full tops assessments)	fitting	regression
		validation	regression
	Douglas-fir	fitting	regression
		validation	regression
Prpn_LIVE	radiata pine	fitting	regression
		validation	regression
	Douglas-fir	fitting	regression
		validation	regression

When considering why random forests have worse performance, note that the regression models of the proportion variables *Tops_prpn_DAM* and *Prpn_LIVE* include *Plot_no* as an observation-level random effect (a specific type of mixed effect) to address over-dispersion in the model response. In comparison, inclusion of *P_stand, P_YOE* or *P_YOM* as a mixed effect addresses data clustering by these stand-level attributes, which occur at a higher level in the data hierarchy than the plot-level response data. While a relatively small number of categories, such as the 85 levels of *P_stand*, the 28 levels of *P_YOE*, and the eight levels of *P_YOM* can be used as potential explanatory variables to the random forest implementation used in this research, attempting to use the 942 levels of *Plot_no* causes the random forest calculation to fail[4]. While there is an attempted formulation of random forests specifically for mixed-effects with variables with high numbers of levels[5], it is quite

[4] These are the levels for the entire dataset. The number of levels by these categories and by species are in Table 2-10 and Table 2-11.

[5] As detailed at https://towardsdatascience.com/mixed-effects-random-forests-6ecbb85cb177

experimental in nature and does not currently exist for the R software used in this thesis. Therefore, the model explanatory power differences between regression and random forest models may be due partly to the random forests not being able to utilise the variable *Plot_no*, which is influential in all the logistic regression models.

The variables included in regression and random forest models, by response variable and species, are presented in Table 4-6 and Table 4-7, below. While there are some similarities beween the variables sets regressions (where the variables are analyst-chosen) and random forests (where the variables are algorithm-chosen), there are also many differences. The most noticeable difference is that, despite creating the random forest models with the *cForest* algorithm, which is intended to handle high levels of correlation in predictor variables, the random forest models generally include several highly correlated variables. This author suspects, but cannot prove, that the random forest models created in this study remain over-reliant on highly correlated variables. This is particularly so when multiple inclusions are made from *P_age_meas* and the weather variables, or multiple inclusions are made from aspect variables (the set *card_4wayN*, *card_4wayNE*, *card_8way*), or multiple inclusions are made from morphometric protection index variables (the set *MPI_100*, *MPI_200*, *MPI_500*, *MPI_1000*, *MPI_2000*).

As outlined in Methods section 2.6.4.4, three differing input variable sets were used during random forest analysis. These were 1) all explanatory variables; 2) the best ten explanatory variables from results of the all-variables model, 3) the same explanatory variables as for the corresponding regression. In all random forest models, variable set 2) gave the best model performance.

Table 4-6: comparison of occurrence of variables in regression and random forest models of radiata pine.

response variable	model type	explanatory variables describing				
		trees	silvicultural history	topography	weather	plots
P_tree_ht_ mean_BRKN	linear regression	P_sph_equiv_cs P_BA_ha_equiv_cs	P_pru_prpn_cs	P_alt_cs	u_wind_tim_cs	P_YOE P_YOM
	random forest	P_age_meas P_tree_ht_mean_NRML P_BA_ha_equiv	P_pru_prpn P_pru_ht		u_rain u_air_pr u_rain_wind_tim u_wind_tim	P_YOM
***Tops_prpn_DAM*, all plots**	logistic regression	Prpn_LIVE_cs	P_thinned	card_4wayN MPI_1000_cs		Plot_no P_YOE P_YOM
	random forest	P_sph_equiv	P_pru_prpn	P_slope P_alt card_4wayN MPI_100 MPI_200		P_YOE P_YOM P_stand
***Tops_prpn_DAM*, plots with full tops assessment**	logistic regression	P_dbh_mean_NRML_cs P_sph_equiv_cs Prpn_LIVE_cs		card_4wayNE MPI_200_cs		Plot_no P_stand
	random forest	P_sph_equiv Prpn_LIVE	P_pru_prpn	P_slope MPI_100 MPI_200 MPI_500 MPI_1000 MPI_2000		P_YOE
Prpn_LIVE	logistic regression	P_sph_equiv_cs Estab_sph_cs Tops_prpn_DAM_cs	P_pru_prpn_cs	P_alt_cs card_4wayNE	u_wind_tim_cs	Plot_no
	random forest	P_age_meas Tops_prpn_DAM	P_pruned P_pru_prpn		u_wind_tim u_air_pr u_rain	P_YOM P_stand

Table 4-7: comparison of occurrence of variables in regression and random forest models of Douglas-fir.

response variable	model type	explanatory variables describing				
		trees	silvicultural history	topography	weather	plots
P_tree_ht_ mean_BRKN	linear regression	P_tree_ht_mean_NRML_cs Tops_prpn_DAM_cs	P_thinned			P_stand
	random forest	P_age_meas P_tree_ht_mean_NRML P_slend_mean	P_thinned	P_alt	u_rain_wind_tim	P_YOE P_YOM P_stand
Tops_prpn_DAM	logistic regression	P_BA_ha_equiv_cs Prpn_LIVE_cs		card_4wayNE		Plot_no P_stand
	random forest	Estab_sph Prpn_LIVE		card_4wayN card_4wayNE card_8way MPI_500 MPI_2000 WindSheltNE1		P_stand P_YOE
Prpn_LIVE	logistic regression	P_sph_equiv_cs P_tree_ht_mean_NRML_cs Tops_prpn_DAM_cs				Plot_no
	random forest	P_sph_equiv P_slend_mean P_tree_ht_mean_NRML Tops_prpn_DAM	P_thinned	P_alt WindSheltNE1	u_min_temp	P_YOM P_YOE

4.7.2 Windspeed and wind direction

An analysis of the wind data for Timaru Aerodrome direction supports a core assumption made in this study, namely that tree damage is caused by wind and snow. Plotting the top two percent of wind speeds at the Timaru Aerodrome[6] by month and direction together, where frequency is proportional to width of the plot figures, shows that the strongest winds do not usually come from directions that are between 0 degrees and 135 degrees (from north clockwise to south-east). This is especially the case in the coldest (and perhaps most snowy) months of June and July.

This accords with the presence of the variable *card_4wayNE* in the model discussed in section 4.3.3, where the cardinal directions north-east and south-east have similar proportions of damaged trees, and the cardinal directions north-west and south-west have proportions of damaged trees that are less than the NE/SE group, and also different to one another. Taken together, these points indicate the effect of lee slope (aspects opposite the wind direction) is real, and that lee slopes at Geraldine can be approximated as being from north clockwise to south-east. Appendix 6.4.1 contains a more extensive analysis of wind records.

[6] i.e. those same records as used to create the variable *u_wind_tim*

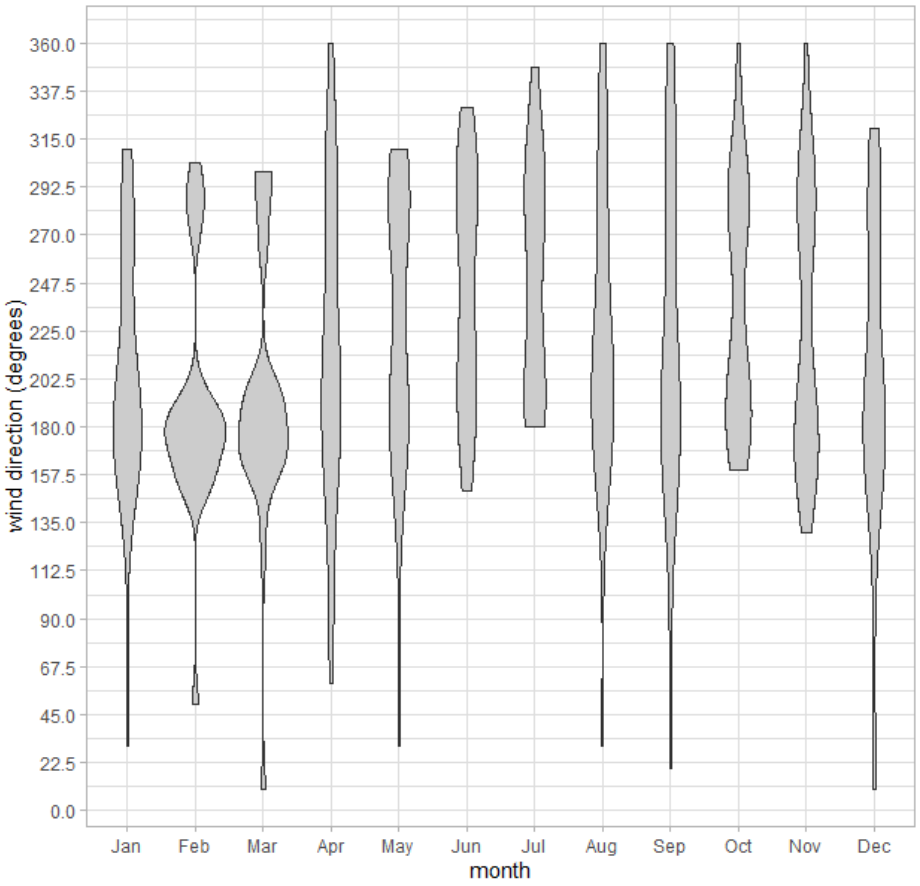

Figure 4-4: distribution of top 2% (29.7 km/hr and above) of daily 9 am wind direction records for Timaru Aerodrome, 31/12/1971 to 31/12/2016, by month and direction.

Ledgard (1982) contented that all windthrow at Geraldine Forest had been due to strong north-westerly winds, which affected mostly very old stands or recently thinned stands. That may have been true for the period cited in that study (1971 – 1981), but seems unlikely over the longer study period here, given the results shown in Figure 4-4, above.

4.7.3 Lack of usefulness of weather variables

During model development of the regression models of radiata pine *P_tree_ht_mean_BRKN* and *Tops_prpn_DAM* (with only plots with known tops included), it became apparent that the weather variables offered no improvement in explantory power relative to the plot age (*P_age_meas*). While the weather variables *u_rain*, *u_rain_wind_tim*, and *u_air_pr* appeared in the random forest model for radiata pine *P_tree_ht_mean_BRKN*, *P_age_meas* was also included, and had much higher relative importance (see Figure 3-3). Therefore, the treatment given to weather variables in this thesis, which was to extract how many 'bad weather days' a plot had experienced, is not a useful technique. Future researchers would do as well with a straightforward plot age measurement, unless their dataset allows specific episodes damage to trees and the causal damaging events to be linked.

4.7.4 Potential alternative response variables not pursued

This study's data suggests that tree top breakage occurs repeatedly in some individual trees. For the radiata pine plots with all tops assessed, there is a weak but definite (0.16) positive correlation between the proportion of tops damaged and the proportion of trees that have a first fork. As trees fork because some influence has removed or killed the terminal bud, this suggests that some previously broken tops have re-grown and left forks in their place. Those forks are then described when the trees are measured. Consideration was given to creating a response variable that included forked tops, broken tops, and windblown trees, instead of the current response variable *Tops_prpn_DAM* that includes broken tops and windblown trees, but it was decided that this response variable did not represent something forest managers want to know, as forked trees do have a functional (if misshapen) top.

Consideration was given in this study to a creating a response variable that comprised trees that had died and were damaged (and so are assumed to have died from the damage), instead of just trees that had died. However, such a variable would have had a low percentage of non-zero values, and as seen from the results chapter (for example, Figure 3-7) the modelling techniques used in this study do not perform well when presented with a high proportion of zeros. Instead, *Tops_prpn_DAM* was included as a candidate explanatory variable for *Prpn_LIVE*, and vice versa.

4.8 Research limitations

Some explanation is required for why this thesis does not deal with the topic of widespread windthrows. As windthrow frequencies in an area increase, at some point the area ceases to be a crop of trees with scattered windthrow and becomes an area of windthrow with scattered standing trees. When windthrow reaches that tipping point, by the judgement of Port Blakely staff, the affected area is mapped out, the stand area is written down, and the affected area is either salvage harvested, cleared and re-planted, or left until the surrounding area is felled, and then prepared for replanting along with the surrounding area. The tree data underpinning this thesis are from ground inventory plots that fell in crops of trees, not in areas that had been mapped out due to windthrow. Port Blakely undertake regular remapping from aerial photos and local knowledge to exclude such areas. Therefore, these research data are from areas that were expected, at the time of inventory, to contain crops of trees.

There are no detailed soil data for Geraldine Forest, and due to time constraints, potentially related variable such as slope position and slope convexity/concavity were not calculated, although the digital elevation model for this study is certainly good enough to support such calculations.

Geraldine Forest has understorey species growing amongst the radiata pine and Douglas-fir that are the subject of this study. It is possible that the understorey species modify the response of radiata pine and Douglas-fir to wind and snow. However, there were no data available regarding understorey in permanent sample plots or inventory plots, which comprise most of the data for this study.

The branch data contained in the plot data available for this study were collected to several different schema, which could not be reconciled to provide meaningful information about branch size or pattern. Therefore, this study does not consider tree branching.

Live crown length, and tree taper, and distance to/direction to nearest clearfell edge were all potentially calculable, but were omitted due to time constraints.

5 Conclusion

This thesis has studied the issue of attritional damage attributed to wind and snow to crops of radiata pine and Douglas-fir at Geraldine Forest, a 5,500 hectare plantation forest in Canterbury, New Zealand. In this forest, approximately 40% of the radiata pine crop suffers from broken tops and scattered windthrow, causing financial losses and operational difficulties. Following interest from the forest managers, Port Blakely, in achieving better understanding of standing-tree breakage and windthrow in these species, a research plan was formulated to investigate these matters.

The initial goals of this research were to utilise data already collected by forest managers or available from the public record, as a basis to: 1) discern whether there is a difference in damage levels between radiata pine and Douglas-fir; 2) identify and understand factors correlating with wind and snow damage to trees at Geraldine Forest; and 3) create empirical models to explain that damage and make sound predictions from new data.

The methods pursued in this study were as follows. First, it was recognised that, with forest measurement plot data and a high-quality digital elevation model available, empirical statistical modelling was possible.

Second, suitable potential explanatory variables were collated, including the identities and locations of the measurement plots, details of the trees in the plots, the silvicultural history of the plots, the topography underlying and surrounding the plots, and the weather history of the plots.

Third, exploratory data analysis revealed which of the potential explanatory variables might relate to the chosen response variables of plot mean broken height (*P_tree_ht_mean_BRKN*), proportion of damaged tree per plot (*Tops_prpn_DAM*), and proportion of live trees per plot (*Prpn_LIVE*).

Fourth, with suggestions from the exploratory data analysis in hand, initial modelling was undertaken to check for differences in damage frequencies between the species of interest. Two model types were decided upon, and applied to all response variables and both species: regression analyses and random forests. The regression analyses included linear and generalised linear models using tree, silviculture, and topographic variables; and linear and generalised linear mixed-effects models that added plot description variables as intercept-only random effects. The random forest models included three types of variable selection: a) allowing the random forest to utilise all explanatory variables, b) restricting the random forest to the ten most important variables from a), and c) requiring the random forest to use the same variables as the best corresponding regression analysis.

The research goals gave rise to a set of research questions, the answers to which constitute the major findings of this research.

1) Do rates of tree damage differ significantly for radiata pine and Douglas-fir at Geraldine Forest?

   Rates of tree damage differ significantly for radiata pine and Douglas-fir. Douglas-fir has higher mean broken heights, lower proportion damaged, and a lower proportion of live trees than radiata pine. The lower proportion alive may be attributable to factors other than damage from wind and snow; but the other two findings likely reflect lower intrinsic vulnerability of Douglas-fir.

2) How well can damage in radiata pine and Douglas-fir be modelled?

Three of the models created have moderate explanatory power (fitting and validation R^2 >= 0.4) and moderate bias (slope of bias check for fitting and validation data >=0.35). These three models are: the linear regression model for radiata pine plot mean broken height (*P_tree_ht_mean_BRKN*); the random forest model for radiata pine plot mean broken height (*P_tree_ht_mean_BRKN*); and the logistic regression model for radiata pine proportion of damaged trees per plot (*Tops_prpn_DAM*), for plots with full tops assessments. Models for other combinations of species, response variable and modelling technique had low explanatory power and high bias.

3) Which modelling approach creates the most explanatory and least biased models?

   Regression models generally had greater explanatory power and less bias than random forest models.

4) Which tree, stand and topographic conditions significantly affect tree damage?

   The linear regression model for radiata pine *P_tree_ht_mean_BRKN* included, for fixed effects, significant positive effects of plot age and plot pruned proportion. North-west aspects had significantly higher broken heights than all other aspects. There was a significant negative effect of plot elevation; the greater the plot's elevation, the lower the main height of breakage. The stand identity was important as a mixed effect, providing an increase in model explanatory power by application of random intercepts.

   The random forest model for radiata pine *P_tree_ht_mean_BRKN* included several effects with above-zero importance. This included a grouping effect on plot mean broken height of plot year of measurement, and positive effects on plot mean broken height of plot age, plot pruned proportion, plot mean pruned height, the mean height of unbroken trees, plot basal area, and a group of weather variables.

   The logistic regression model for radiata pine *Tops_prpn_DAM*, for plots with full tops assessments, included significant positive effects on the per-plot proportion of damaged tops of plot stocking, plot morphometric protection index, and plot mean diameter of unbroken trees. The effects of aspect were variable, with north-west and south-west aspects having significantly lower broken proportions than north-east and south-east aspects. The proportion of trees alive significantly negatively affected the proportion of trees damaged. The stand and plot identities were important mixed effects, providing increases in model explanatory power by application of random intercepts.

5) Can the models developed be used to predict damage from new data?

   Due to the imprecise predictions, the degree of bias and the reliance on mixed-effects models, the three models listed above should not be used to create numeric predictions of damage from new data.

6) Do the research findings suggest forest management practices that may reduce tree damage in radiata pine?

   Taken together, the research findings lead to these suggested management practices to reduce damage to radiata pine at Geraldine Forest:

   - First, consider planting Douglas-fir rather than radiata pine on the most damage-prone topography, which is north-east and south-east aspects, and/or situations with high values of the morphometric protection index.

If radiata pine is planted, then:

- choose a low final stocking: high stocking is associated with higher proportions of damage.
- choose a short rotation: large tree diameters are associated with higher proportions of damage.
- consider pruning, as pruned trees have higher broken heights, leaving more salvageable tree below the break.
- plant radiata pine at lower elevations where growth is faster, allowing a desirable piece size in a shorter rotation.

Returning now to discussing the research goals, we find that the first goal, to discern whether there is a difference in damage levels between radiata pine and Douglas-fir, was accomplished. The second goal, of identifying and understanding factors correlating with damage, was accomplished for three of the species, response variable, and model type combinations.

The third goal, of creating models that explain damage and can make sound predictions from new data, was not accomplished. This is regrettable, as this would have been the outcome of greatest outright usefulness to the forest managers. The three most satisfactory models, as given above, did identify some factors correlated with wind and snow damage to trees at Geraldine Forest, and were able to somewhat predict damage in both the fitting and validation datasets. However, even these models exhibited bias in predictions, and had difficulty predicting ones and zeros. The two most satisfactory regression models also included mixed effects that relied upon the specific identity of stands or plots, further rendering the models unsuitable for making predictions from new data.

This outcome points the way to possible future research into wind and snow damage to even-aged single-species stands of trees, both at Geraldine Forest and more generally. The first avenue that is apparent is the need for modelling techniques that perform well when predicting true zeros and ones in proportional data. An example of such a technique would be some form of data balancing. Models from unbalanced data tend to model the more common occurrence more accurately at the expense of the less common occurrence, which is unhelpful when the less common occurrence (in this case, tree breakage or windthrow) is the phenomenon of interest. If attempting modelling techniques to predict true zeros and ones in similar data, future researchers may note that zero-added binomial models and a manual hurdle model have been tried in this study, without much success. Whatever techniques are tried should not include zero-*inflated* models: all zeros in these data are true, not systemic, zeros.

The second likely avenue is to address the bias in the model results that (probably) arises from model under-specification, that is, the omission of important explanatory variables. Whilst future research may well work with different data, this study did omit several variables suggested by the literature review, due to either time or difficulty of calculation, or an outright lack of data.

Something entirely missing from this study was soil data. The New Zealand soil data are of very low resolution for Geraldine Forest, but perhaps some topographic indices such as slope position or slope convexity/concavity could be beneficial as proxies for potential rooting depth or seasonal waterlogging of soil. Another dataset lacking in this study was properly-kept silvicultural data, which precluded the use of the timing or severity of thinning in models. The effects of the proximity of and direction to recent harvest edges have also been omitted from this study. While other authors (Mitchell et al. (2001), Lanquaye-Opoku and Mitchell (2005)) found a relationship between damage

levels and harvest edges, distances from plots to harvest edges were excessively time-consuming to calculate. Two other variables included by some authors, but omitted in this study due to the time required for calculation, were live crown length or size (for example Wrathall (1989), Scott and Mitchell (2005), and tree taper (for example, Aubrey et al. (2007), Wallentin and Nilsson (2014)), although for tree taper the related height/diameter ratio was included in this study, and did not prove influential. The variable this author most regrets not being able to include, because of previous field observations of trees breaking at whorls of large branches in other New Zealand planation forests, is branch size. Unfortunately, differences among the various types of field survey plots available s made the data branch data irreconcilable.

A technique developed in this study that this author would recommend to future researchers is the treatment of aspect as a four-way or eight-way categorical variable based on predominant aspect of the field measurement plots (see Methods section 2.4.2.4 for a description of this technique). The aspect variables were influential in two of the three most satisfactory models. They were simple to calculate, and avoided the expression of aspect as a deviation from a direction chosen *a priori*.

A technique developed in this study that was ultimately quite unproductive, and so is not recommended for future use, is the extraction of weather data by means of threshold values. Calculating how many very windy days (or rainy, or cold) a plot of trees had seen and using that in models offered very little extra benefit in comparison to the more straightforward plot age variable. This author suspects that if weather data are to be of benefit in modelling damage to trees, it must be in some framework that relates weather data to specific instances of damage. It also is possible that the weather variables calculated regarding wind are unrepresentative for Geraldine Forest, as the nearest long-run wind data records are from Timaru Aerodrome, 38 kilometres away.

Wind damage to plantation forests in New Zealand has been and seems likely to remain a problem for managers of the forests, causing loss of productivity, loss of harvest income, and operational difficulties. For radiata pine, this study has discovered some topographic and stand variables that correlate with damage to trees at Geraldine Forest, and gone some way to predicting damage levels from those variables. This author hopes that other researchers will take up the challenge of predicting wind and snow damage in other New Zealand plantation forests from pre-existing information, and that they will be able to offer management suggestions and damage predictions specific to those forests, by drawing on the concepts in this thesis.

6 Appendices

6.1 Plot data processing

6.1.1 Exclusions

Plots with uncertain locations, for example, those with a field crew note about erratic GPS behaviour, were excluded, as were mirage plots. Only PSPs with co-ordinates for plot location were considered for use: many historic PSP records for Geraldine Forest lack co-ordinates.

Plots containing two or more non-crop stems of woody trees other than the dominant plot were excluded from the data. This was a compromise. The ideal practice would been to exclude every plot containing any non-crop stems of woody trees. However, this would have removed too many plots from the analysis. In plots where a non-crop stem was retained, this was re-classified to the crop species. As any tree occupies growing space and therefore affects its neighbours by competition, reclassification is preferable to arbitrary removal.

Below follows a list of plots that were removed from the analysis set, prior to prior to the plots being split into fitting and validation data, and the reason why they were omitted.

Plot_no	Reason for exclusion
GRLD002002_12_001	Not pure radiata pine, also Douglas-fir
GRLD010001_12_004	Plot downsized in field, but new size not recorded
GRLD020001_11_005	Field crew comment about poor GPS behaviour
GRLD021001_11_011	Plot shifted in field, but shifted location not recorded
GRLD027001_11_001	Location data only, no tree data
GRLD027001_11_002	Location data only, no tree data
GRLD027001_11_003	Location data only, no tree data
GRLD027001_11_004	Location data only, no tree data
GRLD027001_11_005	Location data only, no tree data
GRLD027001_11_006	Location data only, no tree data
GRLD027001_11_007	Location data only, no tree data
GRLD027001_11_008	Location data only, no tree data
GRLD027001_11_009	Location data only, no tree data
GRLD027001_11_010	Location data only, no tree data
GRLD027001_11_011	Location data only, no tree data
GRLD027001_11_012	Location data only, no tree data
GRLD027001_11_013	Location data only, no tree data
GRLD027001_11_014	Location data only, no tree data
GRLD027001_11_015	Location data only, no tree data
GRLD027001_11_016	Location data only, no tree data
GRLD027001_11_017	Location data only, no tree data
GRLD034001_15_003	Overlap with LIDAR_16_097
GRLD034001_15_006	Mirage plot (a technique for handling plots on edges that leads to some trees being measured twice)
GRLD034002_11_002	Overlap with GRLD034002_15_001

Plot_no	Reason for exclusion
GRLD034002_11_009	Overlap with LIDAR_16_163
GRLD101001_13_002	Field crew comment about poor GPS behaviour
GRLD102001_14_003	Excluded for overlap with BP_2_0_3_0_16
GRLD102001_14_005	Overlap with LIDAR_16_198
GRLD103001_13_005	Overlap with GRLD103001_15_014
GRLD103001_13_006	Field crew comment about poor GPS behaviour
GRLD103001_13_010	Overlap with GRLD103001_15_003
GRLD103001_13_014	Overlap with GRLD103001_15_008
GRLD103001_15_001	Plot upsized in field, but new size not recorded
GRLD103001_15_002	Plot upsized in field, but new size not recorded
GRLD103001_15_010	Field crew comment about poor GPS behaviour
GRLD103001_15_015	Mirage plot (a technique for handling plots on edges that leads to some trees being measured twice)
GRLD202003_14_001	Location data only, no tree data
GRLD202003_14_005	Overlap with LIDAR_16_097
GRLD204001_15_005	Field crew comment about poor GPS behaviour
GRLD204001_15_008	Plot downsized in field, but new size not recorded
GRLD307001_13_007	Plot downsized in field, but new size not recorded
GRLD308001_13_026	Field crew comment about poor GPS behaviour
GRLD414001_11_023	Plot shifted in field, but shifted location not recorded
GRLD414001_11_024	Plot shifted in field, but shifted location not recorded
GRLD414001_11_029	Field crew comment about poor GPS behaviour
GRLD414001_11_031	Plot shifted in field, but shifted location not recorded
GRLD416001_11_006	Overlap with LIDAR_16_079
GRLD417001_11_011	Overlap with
GRLD418004_14_010	Overlap with LIDAR_16_156
GRLD504001_15_035	Overlap with LIDAR_16_123
GRLD504001_15_058	Field crew comment about poor GPS behaviour
GRLD504001_15_059	Overlap with LIDAR_16_201
GRLD504001_15_068	Overlap with LIDAR_16_038
GRLD504001_15_071	Overlap with LIDAR_16_016
GRLD507001_15_089	Field crew comment about poor GPS behaviour
GRLD507001_15_118	Mirage plot (a technique for handling plots on edges that leads to some trees being measured twice)
GRLD507002_15_123	Mirage plot (a technique for handling plots on edges that leads to some trees being measured twice)
GRLD507002_15_124	Mirage plot (a technique for handling plots on edges that leads to some trees being measured twice)
GRLD514001_12_007	Overlap with LIDAR_16_049
GRLD515002_12_005	Overlap with LIDAR_16_218
GRLD515002_12_006	Field crew comment about poor GPS behaviour
GRLD515002_12_012	Field crew noted some trees not measured
GRLD517001_12_005	Overlap with LIDAR_16_003
LIDAR_16_002	Field crew comment about poor GPS behaviour
LIDAR_16_023	Location data only, no tree data
LIDAR_16_027	Not pure radiata pine, also Douglas-fir
LIDAR_16_046	Location data only, no tree data
LIDAR_16_088	Mixture of age classes

Plot_no	Reason for exclusion
LIDAR_16_131	Location data only, no tree data
LIDAR_16_162	Plot was all Douglas-fir, the stand is radiata pine, some stand variables would therefore be wrong
LIDAR_16_166	Not pure radiata pine, also Douglas-fir
LIDAR_16_194	Not pure radiata pine, also Douglas-fir
LIDAR_16_215	Location data only, no tree data

6.1.2 Map projection

The New Zealand Transverse Mercator (NZTM) map projection is a square Cartesian grid, measured in metres from an arbitrary origin. It is the current official New Zealand map projection (Land Information New Zealand, 2008), and has been used throughout this research.

6.1.3 Processing plot data

Data was provided in several formats and several map projections. Methods of data processing are described in sections 6.1.3.1 to 6.1.3.3 to below.

6.1.3.1 Inventory plots

Inventory plot location data collected from 2011 to 2014 required conversion from the MapInfo TAB format. These steps were followed:

- The QuickImport tool in ArcGIS was used to convert TAB files to personal geodatabase tables.
- An ArcGIS Model Builder script was used to iterate through the personal geodatabases and write them out to individual point feature classes in an ArcGIS file geodatabase, with the naming convention <original file name>_<year of inventory>.
- I then checked the projection information for all feature classes: projections were inconsistent and some were unknown to ArcGIS, but all the point feature classes included plot co-ordinates as attributes. Some were in NZTM co-ordinates; some were in decimal degrees.
- Therefore, all point feature classes were exported to (non-spatial) file geodatabase tables, then were re-created as fresh NZTM point feature classes using ArcGIS's Make XY Event Layer and Export to Feature class tools.
- Point feature classes were then merged.

Inventory plot location data collected in 2015 had no associated spatial data, but did include plot co-ordinates as attributes. To create spatial data in this case, these steps were followed:

- Plain-text tab delimited files containing the plot names and co-ordinates were created, on a stand-by-stand basis.
- The Make XY Event Layer and Export to Feature class tools were used to created point feature classes
- Point feature classes were then merged, then merged again with the feature classes containing data converted from MapInfo format.

Inventory plot tree data required processing. These steps were followed:

- Inventory plot data was available in Plotsafe format, with one Plotsafe file per stand that had received inventory.

- Each Plotsafe inventory file was exported to .csv format, in a folder structure that matched the folder structure of the provided Plotsafe files.
- The Tree files (one of the output files from the above step) were extracted from the colder structure and combined
- The tree data were manipulated in Microsoft Excel, until they reached the preferred data structure.

6.1.3.2 LiDAR ground plots

- The Create Feature Class tool in ArcGIS was used to create a file geodatabase point feature class from the point shapefile available for the LiDAR inventory.
- The process for compiling LiDAR ground plot tree data was the same as for inventory data.

6.1.3.3 Permanent sample plots (PSPs)

- A plain-text tab delimited file containing the PSP names and co-ordinates was created. PSP original plot names had to be reformatted to include only single underscores, for compatibility with the file geodatabase.
- The Make XY Event Layer and Export to Feature class tools was used to create a point feature class of plot locations in a file geodatabase.
- PSPs are repeat measures of one plot. One measurement that included descriptions of the top status and heights of the trees in the PSP was chosen. If there were several such measurements, the oldest was used. For consistency with inventory and PSP data, only plot measurements after thinning were eligible for inclusion.

6.1.4 Checks on plot location data

The data from this research comes from three types of ground survey – permanent sample plots, inventory plots, and ground plots associated with a LiDAR survey. Inventory and PSP plot locations, were collected with hand-held GPS of unknown quality. The LiDAR ground plots locations, however, were calculated using highly accurate GPS sensors, and the information was also post-processed with ground station readings. Therefore, LiDAR ground plot locations are highly accurate and highly precise, as shown in Figure 6-1 and Figure 6-2, below.

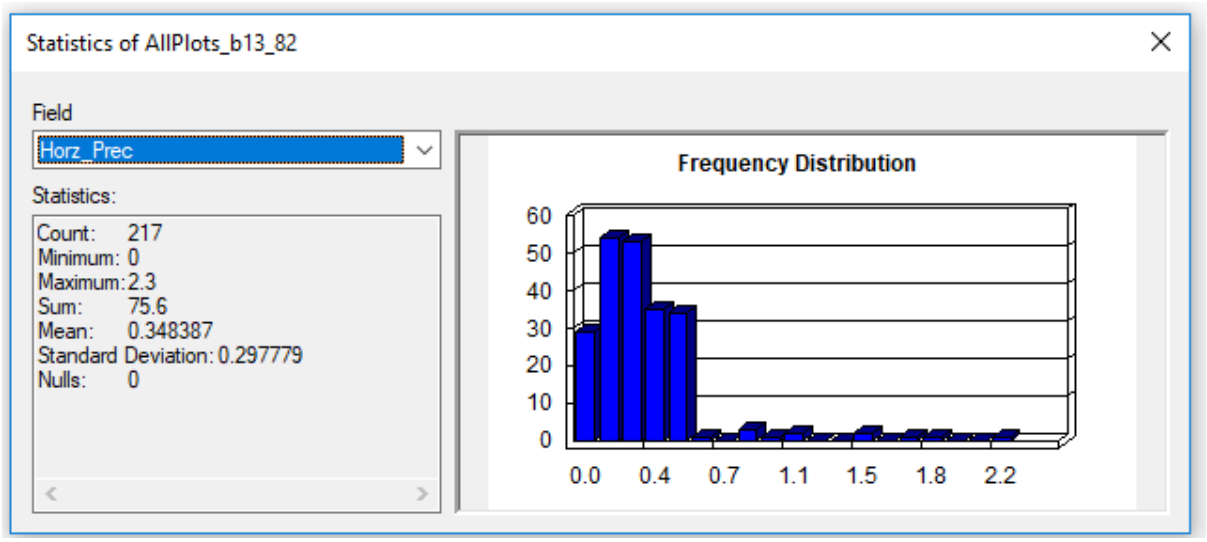

Figure 6-1 Horizontal precision estimates for LiDAR plots.

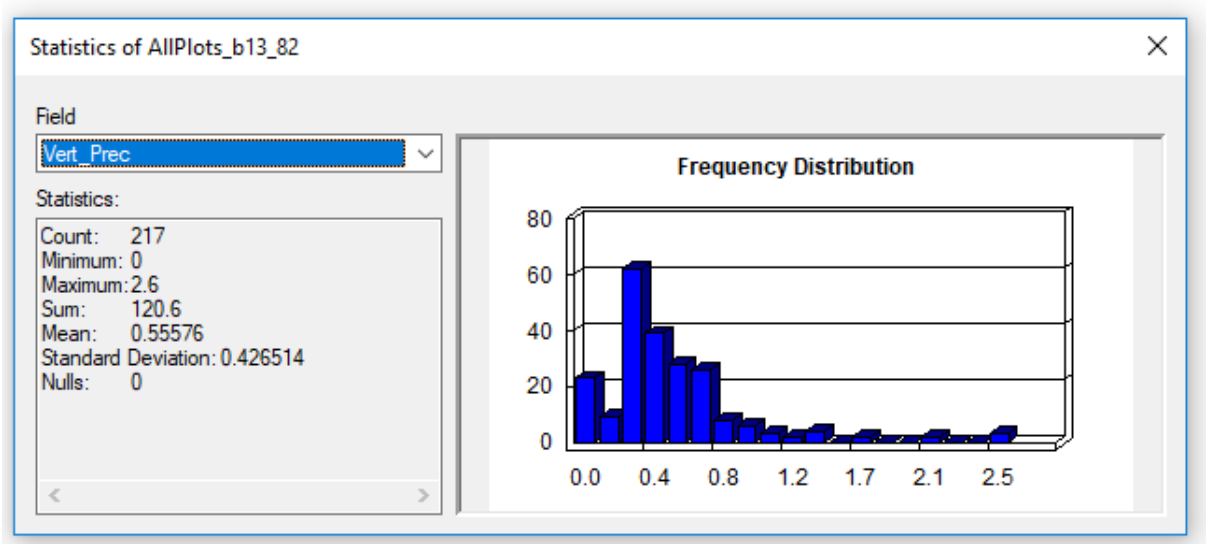

Figure 6-2 Vertical precision estimates for LiDAR plots.

Another check that can be made on the quality of the location data for plots is to compare the observed slope of the plot, as collected by the field crew, with the slope calculated from the intersection of the plot footprint and the slope raster derived from the digital elevation model (see Appendix 6.2.2 for further details). If the plot locations contain significantly more error for inventory plots and PSPs than for LiDAR plots, the differences between observed slope and model slope should be greater, on the whole, for PSP and inventory plots than for LiDAR plots, because the field slope and the terrain model slope will have been are estimated from different areas of ground. The results of this check are shown in Figure 6-3, below, where there appears to be no real difference in slope discrepancies between PSPs and inventory plots, and LiDAR plots.

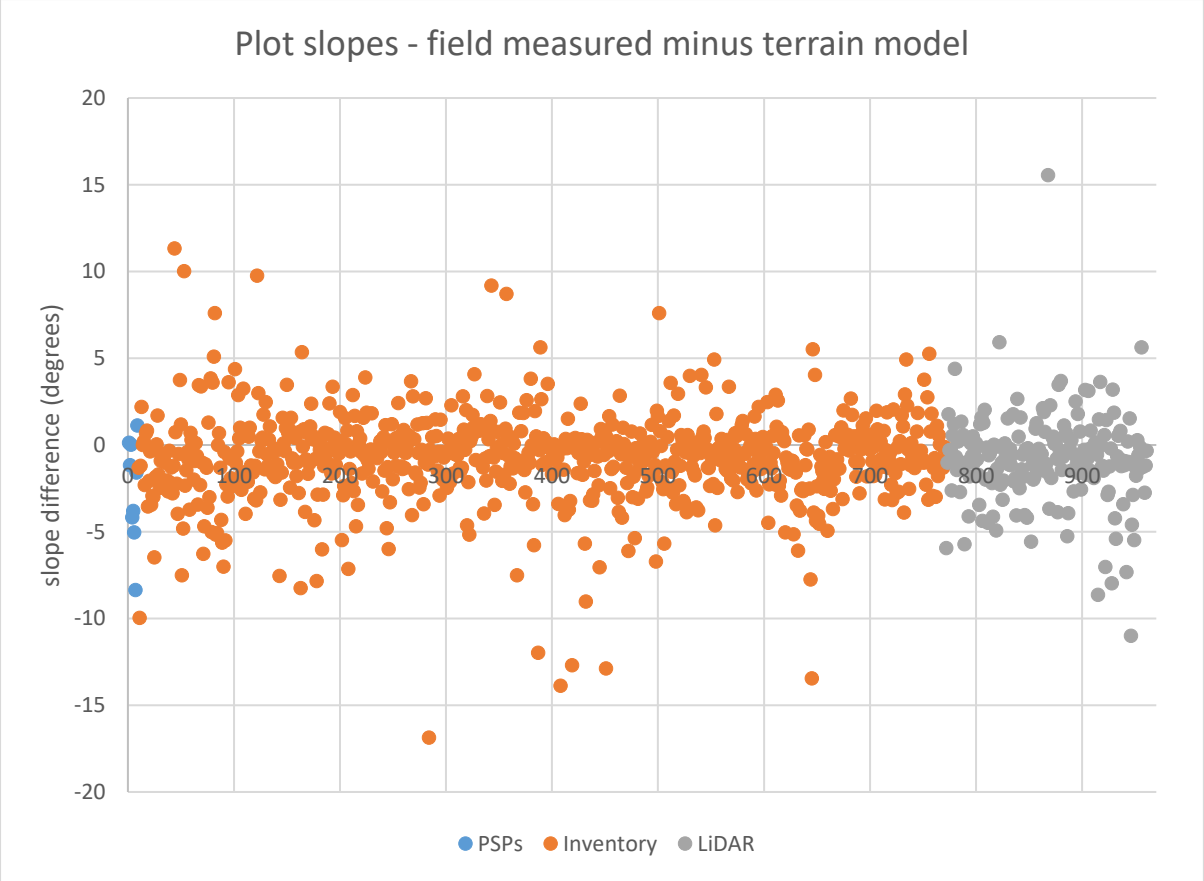

Figure 6-3: reported and DTM-derived plot mean slope, by plot type.

6.1.5 Assumptions associated with silvicultural data

Silvicultural data available for Geraldine Forest were rather incomplete. Some assumptions were made to complete the silvicultural records as far as possible: these are detailed in Table 6-1, below.

Table 6-1: assumptions made to enhance completeness of silvicultural records for Geraldine Forest.

stand or plot	assumption
GRLD010001_12: all plots in stand	final stocking inferred from 1st prune details
GRLD034002_11: all plots in stand	thin stocking inferred from 3rd prune details
GRLD044001_12: all plots in stand	thin stocking inferred from 2nd prune details
GRLD104001_13: all plots in stand	thin stocking inferred from 2nd prune details
GRLD309002_12: all plots in stand	thin stocking inferred from 3rd prune details
GRLD409001: all plots in stand	thin stocking inferred from 3rd prune details
GRLD417001: all plots in stand	thin stocking inferred from 2nd prune details
GRLD418004_14: all plots in stand	thinning inferred from establishment stocking/ current stocking difference
GRLD517001_12: all plots in stand	thin stocking inferred from 2nd prune details
GRLD517002_12: all plots in stand	thin stocking inferred from 2nd prune details
LIDAR_16_005	thin stocking inferred from 3rd prune details
LIDAR_16_009	thinning inferred from establishment stocking/ current stocking difference
LIDAR_16_011	thin stocking inferred from 2nd prune details

stand or plot	assumption
LIDAR_16_015	thin stocking inferred from 2nd prune details
LIDAR_16_021	thinning inferred from establishment stocking/ current stocking difference
LIDAR_16_024	thin stocking inferred from 2nd prune details
LIDAR_16_030	thin stocking inferred from 2nd prune details
LIDAR_16_040	thin stocking inferred from 3rd prune details
LIDAR_16_041	thin stocking inferred from 3rd prune details
LIDAR_16_044	thin stocking inferred from 2nd prune details
LIDAR_16_050	thin stocking inferred from 3rd prune details
LIDAR_16_061	thin stocking inferred from 3rd prune details
LIDAR_16_063	thin stocking inferred from 3rd prune details
LIDAR_16_065	thin stocking inferred from 3rd prune details
LIDAR_16_074	thin stocking inferred from 2nd prune details
LIDAR_16_081	thin stocking inferred from 2nd prune details
LIDAR_16_082	thin stocking inferred from 3rd prune details
LIDAR_16_083	thin stocking inferred from 2nd prune details
LIDAR_16_091	thin stocking inferred from 2nd prune details
LIDAR_16_092	thin stocking inferred from 2nd prune details
LIDAR_16_099	thin stocking inferred from 2nd prune details
LIDAR_16_101	thinning inferred from establishment stocking/ current stocking difference
LIDAR_16_104	thin stocking inferred from 2nd prune details
LIDAR_16_125	thin stocking inferred from 3rd prune details
LIDAR_16_132	final stocking inferred from 1st prune details
LIDAR_16_135	thin stocking inferred from 2nd prune details
LIDAR_16_142	thinning inferred from establishment stocking/ current stocking difference
LIDAR_16_144	thin stocking inferred from 2nd prune details
LIDAR_16_148	thin stocking inferred from 2nd prune details
LIDAR_16_149	thin stocking inferred from 2nd prune details
LIDAR_16_155	thin stocking inferred from 3rd prune details
LIDAR_16_156	thinning inferred from establishment stocking/ current stocking difference
LIDAR_16_163	thin stocking inferred from 3rd prune details
LIDAR_16_171	thin stocking inferred from 2nd prune details
LIDAR_16_175	final stocking inferred from 1st prune details
LIDAR_16_179	thinning inferred from establishment stocking/ current stocking difference
LIDAR_16_180	thin stocking inferred from 3rd prune details
LIDAR_16_185	thin stocking inferred from 2nd prune details
LIDAR_16_187	thin stocking inferred from 3rd prune details
LIDAR_16_189	thin stocking inferred from 3rd prune details
LIDAR_16_203	thin stocking inferred from 2nd prune details
LIDAR_16_204	thin stocking inferred from 2nd prune details
LIDAR_16_211	thin stocking inferred from 2nd prune details
LIDAR_16_219	thin stocking inferred from 2nd prune details

6.2 Topographic variable extraction

6.2.1 Calculating plot footprints

As is customary in forestry, the plot areas in the base data of this research are planar areas. Plots placed on slopes greater than zero degrees (flat) will have had had their radii adjusted upwards at the time of plot installation, to create circular plots whose physical area varies, but whose planar areas are consistent for each survey task. Therefore, each plot point must be buffered by the adjusted radius of the plot, to give the slope-corrected circular area, which is the appropriate footprint for calculating terrain variables by plot.

Unfortunately, although the plot radii are listed in the inventory files in Plotsafe format, they are not exported from the Plotsafe format to the corresponding .csv files. Plot size, on the true map plane, in hectares, and plot slope, in degrees, are exported.

Rather than seek to re-program the data export process, which is embedded in the Plotsafe product, or manually read the data, plot radius was calculated from the slope and area. Given that $\text{radius}_{\text{sloping plot}} = \text{radius}_{\text{planar plot}} / \sqrt{\cos(\text{plot slope})}$, the calculations for a nominal example are given in Table 6-2, below.

Table 6-2: example of calculation of plot footprint size.

plot number	plot area (ha)	plot area (square m)	slope (degrees)	slope (radians)	radius of planar plot (m) r planar = sqrt(plot area square m/π)	radius of sloping plot r sloping = r planar / sqrt(cos(slope$_{radians}$))
XXXXX	0.06	600	28	0.489	13.82	14.71

6.2.2 Calculating slope and elevation for plots

Plot mean elevations and plot mean slopes were calculated using the Zonal Statistics as Table tool in ArcGIS 10.6, using the plot footprints as a vector input and the elevation and slope surfaces provided by Port Blakely as raster inputs. Plot elevation is the mean elevation of all pixels underlying the plot footprint. Plot slope is the mean slope of all pixels underlying the plot footprint.

6.2.3 Calculating aspect for plots

Predominant plot aspect was calculated for three alternative categorisations of aspect. The aspects were calculated in ArcGIS and R. First, the existing aspect raster was reclassified to three alternative rasters, one for each categorisation scheme, with nominal values to represent each aspect. Second, each reclassified raster was converted to polygons. Third, the plot footprint polygons were intersected with the aspect polygons, to yield area of each aspect for each plot. Fourth, the results of the intersection were exported and fifth, they were read into R, where the calculation and assignment of the predominant aspect, based on area, was undertaken.

First, the existing aspect raster was re-classified using the Reclassify tool in ArcGIS 10.6 to give three alternative rasters with nominal numeric values to represent the classifications. Second, each reclassified raster was converted to a vector format with the Raster to Polygon tool in ArcGIS 10.6. Third, the Intersect tool in ArcGIS 10.6 was used to intersect plot footprints with each of the three

resulting polygon layers. Fourth, the attribute table for each of the three sets of intersected footprints, which includes areas for each resultant polygon, was exported to .csv format. Fifth, the three .csv files were read into R, which was used to a) give a string value to represent the cardinal directions equivalent to each nominal raster value, b) sum the area of the plot footprint of each cardinal direction and c) extract the largest of the summed areas and return that as the predominant aspect for the plot. An example of the output obtained is shown in Figure 6-4, below. This approach conceptually matches that of Hansen and Cranson (2016), who also used predominant ('majority', in their parlance) aspect of a plot as its overall aspect.

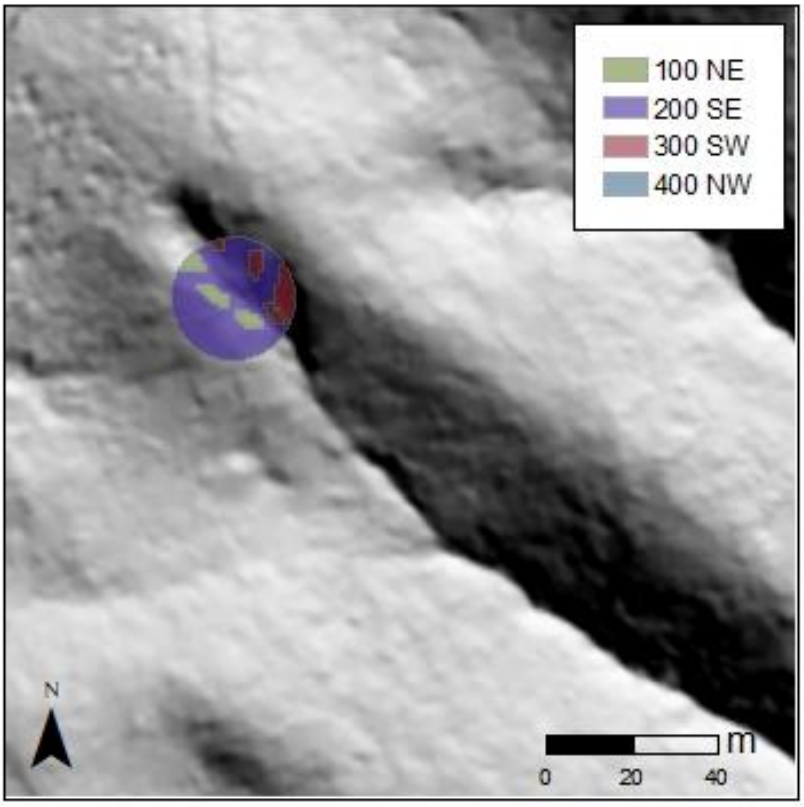

Figure 6-4 Plot GRLD103001_15_009 on the 4-way NE, SE, SW, NW classification, showing the correspondence between nominal raster values in the GIS and categories assigned in R.

Each of the aspect classifications comprised a set of boundaries that divided the compass rose (see Figure 6-5) into evenly-sized groups. The four-way classification of aspect to north, east, south, and west has these characteristics:

- North: 315 to 360 degrees, and 0 to 45 degrees, given raster value *10*, given cardinal value *N*
- East: 45 degrees to 135 degrees, given raster value *20*, given cardinal value *E*
- South: 135 degrees to 225 degrees, given raster value *30*, given cardinal value *S*
- West: 225 degrees to 315 degrees, given raster value *40*, given cardinal value *W*

The four-way classification aspect to north-east, south-east, south-west, and north-west has these characteristics:

- North-east: 0 to 90 degrees, given raster value *100*, given cardinal value *NE*
- South-east: 90 degrees to 180 degrees, given raster value *200*, given cardinal value *SE*

- South-west: 180 degrees to 270 degrees, given raster value *300*, given cardinal value *SW*
- North-west: 270 degrees to 360 degrees, given raster value *400*, given cardinal value *NW*

The eight-way classification of aspect has these characteristics:

- North: 337.5 degrees to 360 degrees, and 0 degrees to 22.5 degrees, given raster value *1*, given cardinal value *n*
- North-east: 22.5 degrees to 67.5 degrees, given raster value *2*, given cardinal value *ne*
- East: 67.5 degrees to 112.5 degrees, given raster value *3*, given cardinal value *e*
- South-east: 112.5 degrees to 157.5 degrees, given raster value *4*, given cardinal value *se*
- South: 157.5 degrees to 202.5 degrees, given raster value *5*, given cardinal value *s*
- South-west: 202.5 degrees 247.5 degrees, given raster value *6*, given cardinal value *se*
- West: 247.5 degrees to 292.5 degrees, given raster value *7*, given cardinal value *w*
- North-west: 292.5 degrees to 337.5 degrees, given raster value *8*, given cardinal value *nw*

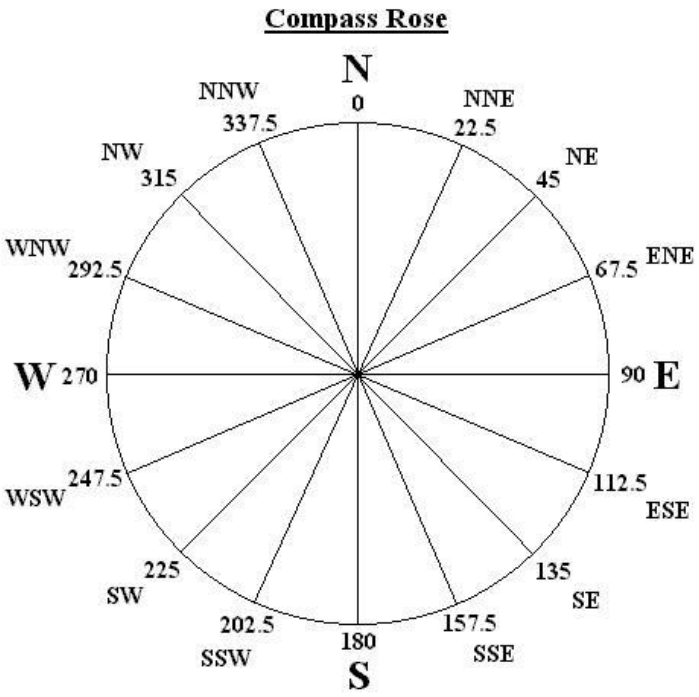

Figure 6-5 Compass rose showing degrees and cardinal directions.
Reproduced from http://www.ke4nyv.com/navigation.htm, with permission of the site owner

6.2.4 Calculating average morphometric protection index (MPI) for plots

Plot MPI values were calculated using the Zonal Statistics as Table tool in ArcGIS 10.6, using the plot footprints as a vector input and the MPI surfaces calculated in SAGA as the raster inputs. Plot MPI is the mean MPI of all pixels underlying the plot footprint.

6.2.5 Variability in the topographic variables

6.2.5.1 Elevation and slope

In this thesis, plot mean elevation and plot mean slope have been used as modelling inputs. Because the location of individual trees within a plot is not known, all topographic variables apply at the plot level. However, summary statistics about the variability of elevation and slope within plots are presented in Table 6-3, along with frequency distributions in Figure 6-6 and Figure 6-7, below.

Table 6-3: minima, maxima and means for per-plot elevation and slope data.

statistic	minimum	maximum	mean
range of elevation within a plot	0.9 m	29.2 m	12.1 m
standard deviation of elevation within a plot	0.1 m	7.7 m	3.0 m
range of slope within a plot	7°	70°	25°
standard deviation of slope within a plot	1.4°	18.2°	4.5 °

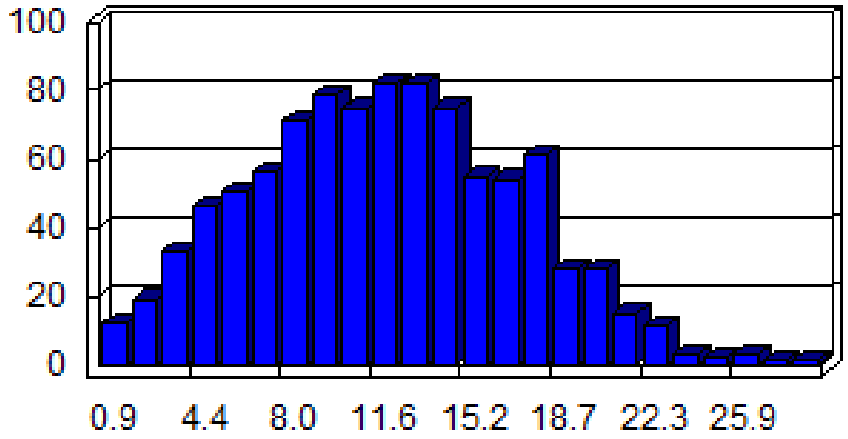

Figure 6-6 Frequency distribution of the range of altitude within a plot.

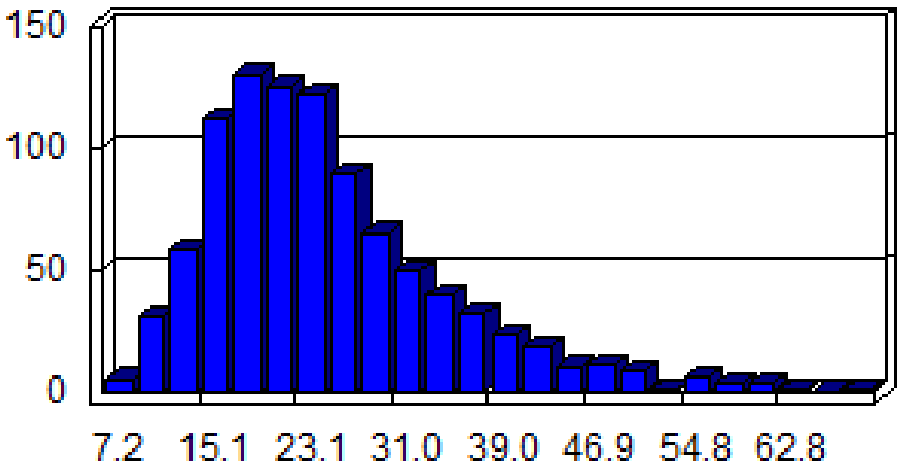

Figure 6-7 Frequency distribution of the range of slopes within a plot.

6.2.5.2 Aspect

As detailed in Methods, the procedure used in this research assigns as the predominant aspect of a plot the aspect that has the largest area within that plot. This means that a plot could be assigned an aspect on the basis of just over 25% of its area, in the case of the four-way classifications, or on the basis of just over 12.5% of its area, in the case of the eight-way classification. Figure 6-8, Figure 6-9 and Figure 6-10 , below, demonstrate that most plots have one predominant aspect, especially when considering the four-way classifications of aspect.

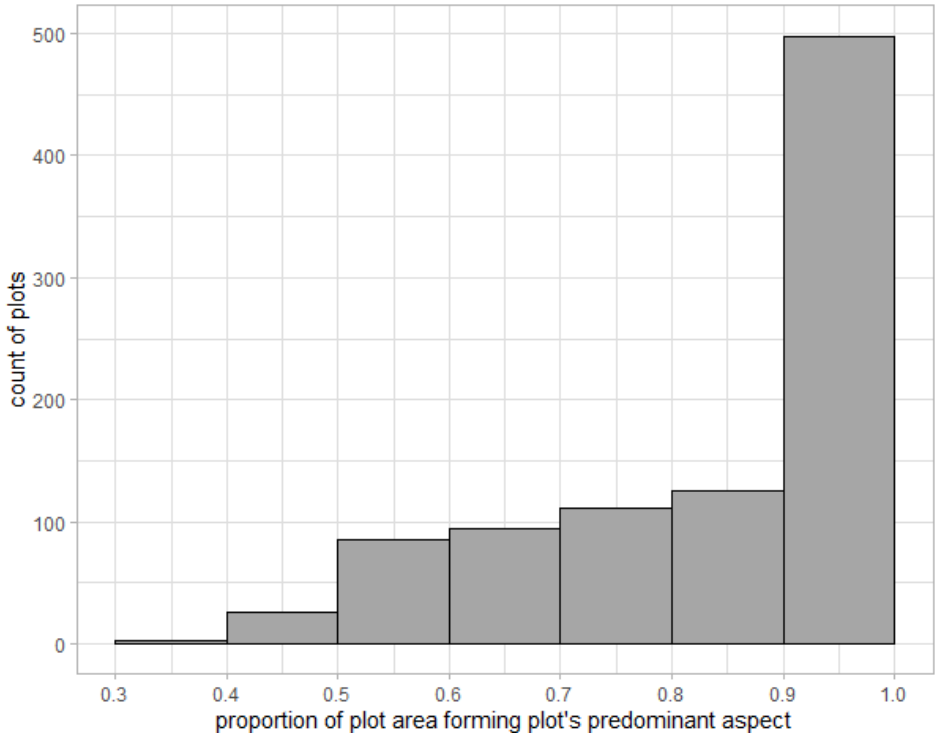

Figure 6-8: analysis of plot predominant aspect, N/E/S/W classification.

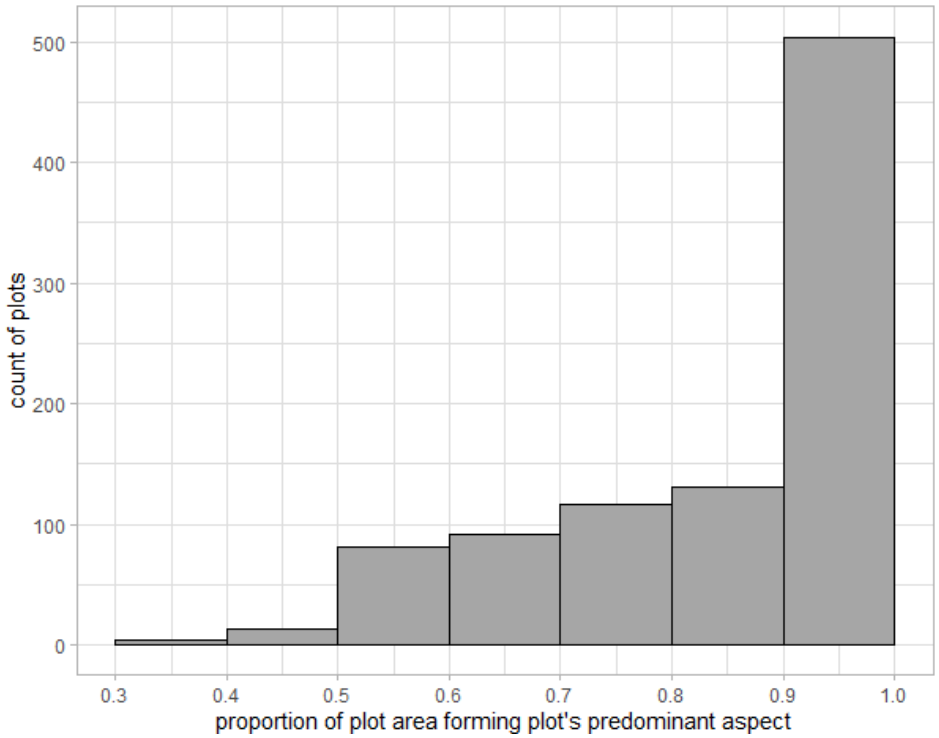

Figure 6-9: analysis of plot predominant aspect, NE/SE/SW/NW classification.

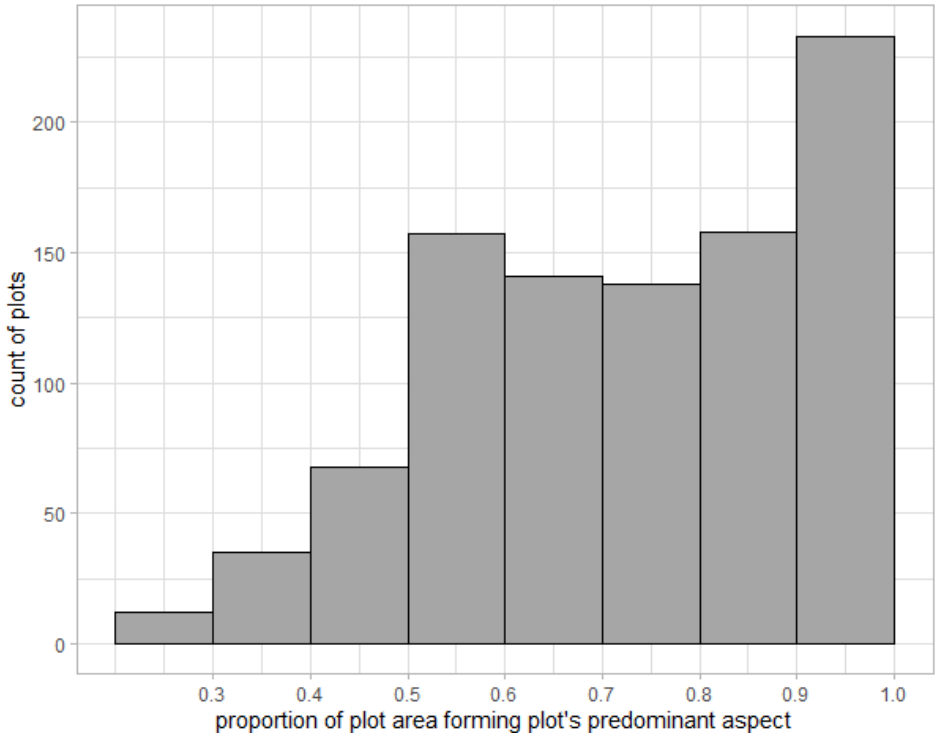

Figure 6-10: analysis of plot predominant aspect, n/ne/e/se/s/sw/w/nw classification.

6.3 Fitting and validation plot identities

Table 6-4 to Table 6-7, below, give the identities of the plots comprising the fitting and validation datasets.

Table 6-4: plot numbers of fitting plots, radiata pine.

BP_2_0_1_0_15	LIDAR_16_075	LIDAR_16_136	LIDAR_16_190
BP_2_0_2_0_15	LIDAR_16_076	LIDAR_16_139	LIDAR_16_191
BP_2_0_3_0_16	LIDAR_16_077	LIDAR_16_140	LIDAR_16_192
CY_560_3_10_0_06	LIDAR_16_080	LIDAR_16_141	LIDAR_16_193
CY_560_3_11_0_13	LIDAR_16_082	LIDAR_16_144	LIDAR_16_198
CY_560_3_6_0_13	LIDAR_16_084	LIDAR_16_145	LIDAR_16_199
CY_560_3_8_0_13	GRLD034001_15_009	LIDAR_16_146	LIDAR_16_200
CY_560_3_9_0_03	GRLD034002_11_001	LIDAR_16_149	LIDAR_16_201
GRLD002002_12_004	GRLD034002_11_003	LIDAR_16_150	LIDAR_16_202
GRLD002002_12_006	GRLD034002_11_004	LIDAR_16_151	LIDAR_16_203
GRLD002002_12_007	GRLD034002_11_006	LIDAR_16_152	LIDAR_16_204
GRLD010001_12_001	GRLD034002_11_007	LIDAR_16_154	LIDAR_16_205
GRLD010001_12_002	GRLD034002_11_008	LIDAR_16_155	LIDAR_16_206
GRLD010001_12_003	GRLD034002_15_001	GRLD202003_14_002	LIDAR_16_207
GRLD010001_12_005	GRLD034002_15_003	GRLD202003_14_003	LIDAR_16_208
GRLD010001_12_006	GRLD034002_15_004	GRLD202003_14_004	LIDAR_16_209
GRLD019001_11_001	GRLD034002_15_005	GRLD202003_14_007	LIDAR_16_210
GRLD019001_11_002	GRLD034002_15_006	GRLD202003_14_008	LIDAR_16_211
GRLD019001_11_003	GRLD034002_15_008	GRLD202003_14_009	LIDAR_16_212
GRLD019001_11_005	GRLD034002_15_009	GRLD202003_14_010	LIDAR_16_214
GRLD019001_11_006	GRLD042001_12_001	GRLD204001_15_002	GRLD410001_11_008
GRLD019001_11_007	GRLD042001_12_002	GRLD204001_15_003	GRLD410001_11_009
GRLD019001_11_008	GRLD042001_12_004	GRLD204001_15_004	GRLD410001_11_010
GRLD019001_11_010	GRLD042001_12_006	GRLD204001_15_006	GRLD410001_11_011
GRLD019001_11_011	GRLD044001_12_001	GRLD204001_15_007	GRLD411001_11_001
GRLD019001_11_012	GRLD044001_12_002	GRLD205002_15_001	GRLD411001_11_003
GRLD019001_11_013	GRLD044001_12_003	GRLD205002_15_002	GRLD411001_11_004
GRLD019001_11_014	GRLD044001_12_004	GRLD205002_15_004	GRLD411001_11_005
GRLD023001_12_001	GRLD044001_12_005	GRLD205002_15_005	GRLD411001_11_006
GRLD023001_12_002	GRLD044001_12_006	GRLD205002_15_006	GRLD411001_11_007
GRLD023001_12_003	GRLD044001_12_007	GRLD205002_15_007	GRLD412001_11_001
GRLD023001_12_004	GRLD101001_13_001	GRLD205002_15_008	GRLD412001_11_002
GRLD023001_12_005	GRLD101001_13_003	GRLD205002_15_009	GRLD412001_11_003
GRLD033003_15_001	GRLD101001_13_004	GRLD205002_15_010	GRLD412001_11_004
GRLD033003_15_002	GRLD101001_13_005	GRLD205002_15_011	GRLD412001_11_005
GRLD033003_15_003	GRLD101001_13_006	GRLD205002_15_012	GRLD412001_11_006
GRLD033003_15_004	GRLD101001_13_007	GRLD205002_15_013	GRLD412001_11_008
GRLD033003_15_005	GRLD101001_13_008	GRLD205002_15_014	GRLD412001_11_011
GRLD033003_15_006	GRLD102001_14_001	GRLD309001_12_001	GRLD414001_11_001

GRLD034001_15_001	GRLD102001_14_004	GRLD309002_12_001	GRLD414001_11_002
GRLD034001_15_002	GRLD102001_14_007	GRLD309002_12_002	GRLD414001_11_003
GRLD034001_15_004	GRLD102001_14_008	GRLD309002_12_003	GRLD414001_11_004
GRLD034001_15_005	GRLD104001_13_001	GRLD409001_13_016	GRLD414001_11_006
GRLD034001_15_007	GRLD104001_13_002	GRLD409001_13_018	GRLD414001_11_008
GRLD034001_15_008	GRLD104001_13_003	GRLD409001_13_019	GRLD414001_11_009
GRLD416001_11_007	GRLD104001_13_004	GRLD409001_13_021	GRLD414001_11_010
GRLD416001_11_008	GRLD104001_13_005	GRLD409001_13_031	GRLD414001_11_011
GRLD416001_11_009	GRLD104001_13_006	GRLD409001_13_032	GRLD414001_11_012
GRLD417001_11_001	GRLD104001_13_007	GRLD409001_13_035	GRLD414001_11_013
GRLD417001_11_003	GRLD104001_13_008	GRLD409001_13_036	GRLD414001_11_014
GRLD417001_11_006	GRLD104001_13_009	GRLD409001_13_037	GRLD414001_11_015
GRLD417001_11_007	GRLD504001_15_069	GRLD409001_13_038	GRLD414001_11_016
GRLD417001_11_008	GRLD504001_15_070	GRLD409001_13_039	GRLD414001_11_017
GRLD417001_11_009	GRLD504001_15_072	GRLD410001_11_001	GRLD414001_11_019
GRLD417001_11_010	GRLD504001_15_073	GRLD410001_11_002	GRLD414001_11_020
GRLD417001_11_012	GRLD504001_15_074	GRLD410001_11_005	GRLD414001_11_022
GRLD418004_14_001	GRLD504001_15_075	GRLD410001_11_006	GRLD414001_11_025
GRLD418004_14_003	GRLD507001_15_076	GRLD410001_11_007	GRLD414001_11_026
GRLD418004_14_005	GRLD507001_15_077	GRLD512001_11_001	GRLD414001_11_027
GRLD418004_14_006	GRLD507001_15_078	GRLD512001_11_002	GRLD414001_11_028
GRLD418004_14_008	GRLD507001_15_079	GRLD512001_11_003	GRLD414001_11_030
GRLD418004_14_009	GRLD507001_15_080	GRLD512001_11_004	GRLD416001_11_001
GRLD504001_15_032	GRLD507001_15_081	GRLD512001_11_005	GRLD416001_11_002
GRLD504001_15_034	GRLD507001_15_082	GRLD512001_11_006	GRLD416001_11_004
GRLD504001_15_036	GRLD507001_15_083	GRLD512001_11_007	GRLD416001_11_005
GRLD504001_15_037	GRLD507001_15_085	GRLD512001_11_008	GRLD517001_12_006
GRLD504001_15_038	GRLD507001_15_087	GRLD512001_11_009	GRLD517001_12_007
GRLD504001_15_039	GRLD507001_15_088	GRLD512001_11_010	GRLD517001_12_009
GRLD504001_15_040	GRLD507001_15_091	GRLD514001_12_001	GRLD517001_12_010
GRLD504001_15_041	GRLD507001_15_092	GRLD514001_12_002	GRLD517001_12_011
GRLD504001_15_042	GRLD507001_15_093	GRLD514001_12_003	GRLD517001_12_012
GRLD504001_15_043	GRLD507001_15_095	GRLD514001_12_004	GRLD517001_12_013
GRLD504001_15_044	GRLD507001_15_096	GRLD514001_12_005	GRLD517002_12_001
GRLD504001_15_045	GRLD507001_15_097	GRLD514001_12_006	GRLD517002_12_002
GRLD504001_15_047	GRLD507001_15_098	GRLD514001_12_008	GRLD517002_12_003
GRLD504001_15_048	GRLD507001_15_099	GRLD514001_12_009	GRLD517002_12_004
GRLD504001_15_049	GRLD507001_15_100	GRLD514001_12_010	GRLD517002_12_005
GRLD504001_15_050	GRLD507001_15_101	GRLD514001_12_011	GRLD517002_12_006
GRLD504001_15_051	GRLD507001_15_102	GRLD514001_12_012	GRLD517002_12_007
GRLD504001_15_052	GRLD507001_15_103	GRLD514001_12_013	GRLD517002_12_009
GRLD504001_15_054	GRLD507001_15_105	GRLD514001_12_014	GRLD517002_12_011
GRLD504001_15_055	GRLD507001_15_106	GRLD514001_12_015	GRLD517002_12_012
GRLD504001_15_056	GRLD507001_15_107	GRLD514001_12_016	GRLD517003_12_001
GRLD504001_15_057	GRLD507001_15_108	GRLD514001_12_017	GRLD517003_12_002

GRLD504001_15_060	GRLD507001_15_109	GRLD514001_12_018	GRLD517003_12_003
GRLD504001_15_061	GRLD507001_15_110	GRLD514001_12_019	GRLD517003_12_005
GRLD504001_15_062	GRLD507001_15_111	GRLD514001_12_020	GRLD517003_12_006
GRLD504001_15_063	GRLD507001_15_112	GRLD515002_12_001	GRLD517003_12_007
GRLD504001_15_065	GRLD507001_15_113	GRLD515002_12_003	GRLD519003_12_001
GRLD504001_15_067	GRLD507001_15_114	GRLD515002_12_004	GRLD519003_12_003
LIDAR_16_026	GRLD507001_15_116	GRLD515002_12_007	GRLD519003_12_004
LIDAR_16_029	GRLD507001_15_117	GRLD515002_12_008	GRLD519003_12_005
LIDAR_16_030	GRLD507001_15_119	GRLD515002_12_009	GRLD519003_12_006
LIDAR_16_031	GRLD507002_15_121	GRLD515002_12_010	LIDAR_16_001
LIDAR_16_033	GRLD507002_15_126	GRLD515002_12_011	LIDAR_16_005
LIDAR_16_034	GRLD507002_15_127	GRLD515002_12_013	LIDAR_16_006
LIDAR_16_035	LIDAR_16_086	GRLD515002_12_016	LIDAR_16_007
LIDAR_16_037	LIDAR_16_087	GRLD515002_12_019	LIDAR_16_008
LIDAR_16_038	LIDAR_16_089	GRLD515002_12_020	LIDAR_16_009
LIDAR_16_040	LIDAR_16_090	GRLD515002_12_021	LIDAR_16_010
LIDAR_16_041	LIDAR_16_091	GRLD516001_12_002	LIDAR_16_011
LIDAR_16_042	LIDAR_16_093	GRLD516001_12_004	LIDAR_16_013
LIDAR_16_043	LIDAR_16_094	GRLD516001_12_005	LIDAR_16_014
LIDAR_16_044	LIDAR_16_095	LIDAR_16_156	LIDAR_16_015
LIDAR_16_045	LIDAR_16_097	LIDAR_16_158	LIDAR_16_016
LIDAR_16_047	LIDAR_16_098	LIDAR_16_159	LIDAR_16_017
LIDAR_16_048	LIDAR_16_099	LIDAR_16_160	LIDAR_16_018
LIDAR_16_050	LIDAR_16_100	LIDAR_16_161	
LIDAR_16_051	LIDAR_16_101	LIDAR_16_164	
LIDAR_16_052	LIDAR_16_102	LIDAR_16_167	
LIDAR_16_053	LIDAR_16_103	LIDAR_16_168	
LIDAR_16_054	LIDAR_16_104	LIDAR_16_169	
LIDAR_16_056	LIDAR_16_108	LIDAR_16_170	
LIDAR_16_057	LIDAR_16_109	LIDAR_16_171	
LIDAR_16_058	LIDAR_16_110	LIDAR_16_172	
LIDAR_16_060	LIDAR_16_111	LIDAR_16_173	
LIDAR_16_061	LIDAR_16_112	LIDAR_16_174	
LIDAR_16_063	LIDAR_16_113	LIDAR_16_175	
LIDAR_16_064	LIDAR_16_115	LIDAR_16_176	
LIDAR_16_065	LIDAR_16_117	LIDAR_16_177	
LIDAR_16_066	LIDAR_16_118	LIDAR_16_178	
LIDAR_16_067	LIDAR_16_119	LIDAR_16_179	
LIDAR_16_068	LIDAR_16_123	LIDAR_16_182	
LIDAR_16_069	LIDAR_16_124	LIDAR_16_183	
LIDAR_16_071	LIDAR_16_125	LIDAR_16_185	
LIDAR_16_072	LIDAR_16_128	LIDAR_16_186	
LIDAR_16_073	LIDAR_16_130	LIDAR_16_187	
LIDAR_16_074	LIDAR_16_132	LIDAR_16_189	

Table 6-5: plot numbers of validation plots, radiata pine.

GRLD002002_12_002	GRLD418004_14_004	LIDAR_16_055	
GRLD002002_12_003	GRLD418004_14_007	LIDAR_16_059	
GRLD002002_12_005	GRLD504001_15_033	LIDAR_16_070	
GRLD019001_11_004	GRLD504001_15_046	LIDAR_16_079	
GRLD019001_11_009	GRLD504001_15_053	LIDAR_16_081	
GRLD019001_11_015	GRLD504001_15_064	LIDAR_16_092	
GRLD034002_11_005	GRLD504001_15_066	LIDAR_16_096	
GRLD034002_15_002	GRLD507001_15_084	LIDAR_16_105	
GRLD034002_15_007	GRLD507001_15_086	LIDAR_16_114	
GRLD042001_12_003	GRLD507001_15_090	LIDAR_16_120	
GRLD042001_12_005	GRLD507001_15_094	LIDAR_16_121	
GRLD044001_12_008	GRLD507001_15_104	LIDAR_16_122	
GRLD102001_14_002	GRLD507001_15_115	LIDAR_16_127	
GRLD102001_14_006	GRLD507002_15_120	LIDAR_16_129	
GRLD202003_14_006	GRLD507002_15_122	LIDAR_16_134	
GRLD204001_15_001	GRLD507002_15_125	LIDAR_16_135	
GRLD204001_15_009	GRLD507002_15_128	LIDAR_16_137	
GRLD205002_15_003	GRLD515002_12_002	LIDAR_16_142	
GRLD205002_15_015	GRLD515002_12_014	LIDAR_16_143	
GRLD309001_12_002	GRLD515002_12_015	LIDAR_16_148	
GRLD309001_12_003	GRLD515002_12_017	LIDAR_16_157	
GRLD409001_13_017	GRLD515002_12_018	LIDAR_16_163	
GRLD409001_13_022	GRLD516001_12_001	LIDAR_16_165	
GRLD410001_11_003	GRLD516001_12_003	LIDAR_16_180	
GRLD410001_11_004	GRLD516001_12_011	LIDAR_16_195	
GRLD411001_11_002	GRLD516001_12_017	LIDAR_16_196	
GRLD411001_11_008	GRLD517001_12_002	LIDAR_16_197	
GRLD412001_11_007	GRLD517001_12_004	LIDAR_16_213	
GRLD412001_11_009	GRLD517001_12_008	SR_3011_0_1_0_15	
GRLD412001_11_010	GRLD517001_12_014		
GRLD414001_11_005	GRLD517002_12_008		
GRLD414001_11_007	GRLD517002_12_010		
GRLD414001_11_018	GRLD517003_12_004		
GRLD414001_11_021	GRLD519003_12_002		
GRLD416001_11_003	LIDAR_16_003		
GRLD416001_11_010	LIDAR_16_012		
GRLD417001_11_002	LIDAR_16_021		
GRLD417001_11_004	LIDAR_16_024		
GRLD417001_11_005	LIDAR_16_032		
GRLD418004_14_002	LIDAR_16_039		

Table 6-6: plot numbers of fitting plots, Douglas-fir.

GRLD020001_11_002	GRLD043001_11_001	GRLD105002_15_031	GRLD208001_13_020
GRLD020001_11_003	GRLD043001_11_003	GRLD110001_13_001	GRLD208001_13_021
GRLD020001_11_004	GRLD043001_11_004	GRLD110001_13_002	GRLD208001_13_022
GRLD020001_11_006	GRLD043001_11_005	GRLD110001_13_003	GRLD208001_13_023
GRLD020001_11_008	GRLD043001_11_006	GRLD110001_13_006	GRLD208001_13_024
GRLD020001_11_009	GRLD043001_11_007	GRLD111002_13_002	GRLD208001_13_026
GRLD020001_11_010	GRLD043001_11_008	GRLD111002_13_003	GRLD208001_13_027
GRLD020001_11_011	GRLD043001_11_009	GRLD111002_13_004	GRLD208001_13_028
GRLD020001_11_012	GRLD043001_11_010	GRLD111002_13_006	GRLD208001_13_029
GRLD020001_11_015	GRLD103001_13_001	GRLD111002_13_007	GRLD208001_13_030
GRLD020001_11_016	GRLD103001_13_002	GRLD111002_13_008	GRLD208001_13_031
GRLD020001_11_017	GRLD103001_13_003	GRLD111002_13_010	GRLD208001_13_032
GRLD020001_11_018	GRLD103001_13_004	GRLD111002_13_013	GRLD209001_12_001
GRLD020001_11_019	GRLD103001_13_007	GRLD111002_13_014	GRLD209001_12_002
GRLD021001_11_001	GRLD103001_13_008	GRLD207002_13_001	GRLD209001_12_003
GRLD021001_11_004	GRLD103001_13_009	GRLD207002_13_002	GRLD209001_12_004
GRLD021001_11_005	GRLD103001_13_012	GRLD207002_13_004	GRLD209001_12_006
GRLD021001_11_006	GRLD103001_13_013	GRLD207002_13_005	GRLD209001_12_007
GRLD021001_11_007	GRLD103001_15_003	GRLD207002_13_006	GRLD209001_12_008
GRLD021001_11_009	GRLD103001_15_004	GRLD207002_13_007	GRLD209001_12_009
GRLD021001_11_010	GRLD103001_15_005	GRLD207002_13_008	GRLD209001_12_010
GRLD021001_11_012	GRLD103001_15_006	GRLD207002_13_009	GRLD209001_12_011
GRLD021001_11_013	GRLD103001_15_007	GRLD207002_13_011	GRLD209001_12_012
GRLD026001_11_001	GRLD103001_15_008	GRLD208001_13_001	GRLD209001_12_013
GRLD026001_11_003	GRLD103001_15_009	GRLD208001_13_002	GRLD209001_12_014
GRLD026001_11_004	GRLD103001_15_011	GRLD208001_13_003	GRLD209001_12_015
GRLD026001_11_005	GRLD103001_15_013	GRLD208001_13_004	GRLD209001_12_016
GRLD026001_11_006	GRLD103001_15_014	GRLD208001_13_005	GRLD209001_12_017
GRLD026001_11_007	GRLD103001_15_016	GRLD208001_13_006	GRLD209001_12_018
GRLD026001_11_009	GRLD105001_15_018	GRLD208001_13_007	GRLD209001_12_019
GRLD026001_11_010	GRLD105001_15_019	GRLD208001_13_009	GRLD209001_12_022
GRLD026001_11_012	GRLD105001_15_020	GRLD208001_13_010	GRLD209001_12_023
GRLD026001_11_014	GRLD105002_15_021	GRLD208001_13_011	GRLD209001_12_024
GRLD026001_11_015	GRLD105002_15_022	GRLD208001_13_012	GRLD209001_12_025
GRLD026001_11_016	GRLD105002_15_023	GRLD208001_13_013	GRLD209001_12_026
GRLD026001_11_017	GRLD105002_15_024	GRLD208001_13_014	GRLD209001_12_028
GRLD026001_11_018	GRLD105002_15_025	GRLD208001_13_015	GRLD209001_12_029
GRLD026001_11_019	GRLD105002_15_026	GRLD208001_13_016	GRLD209001_12_030
GRLD026001_11_020	GRLD105002_15_027	GRLD208001_13_017	GRLD209001_12_031
GRLD026001_11_021	GRLD105002_15_029	GRLD208001_13_018	GRLD209002_13_002
GRLD033001_15_011	GRLD105002_15_030	GRLD208001_13_019	GRLD209002_13_003

GRLD209002_13_004	GRLD307001_13_028	GRLD308001_13_030	
GRLD209002_13_005	GRLD307001_13_029	GRLD308001_13_031	
GRLD209002_13_006	GRLD307001_13_030	GRLD308001_13_032	
GRLD209002_13_007	GRLD307001_13_031	GRLD308001_13_033	
GRLD209002_13_008	GRLD307001_13_032	GRLD308001_13_035	
GRLD209002_13_009	GRLD307001_13_033	GRLD308001_13_036	
GRLD209002_13_010	GRLD307001_13_035	GRLD308001_13_038	
GRLD209002_13_012	GRLD307001_13_037	GRLD308001_13_039	
GRLD209002_13_013	GRLD307001_13_039	GRLD308001_13_040	
GRLD209002_13_014	GRLD307001_13_041	GRLD308001_13_041	
GRLD209002_13_015	GRLD307001_13_042	GRLD308001_13_042	
GRLD304002_13_001	GRLD307003_13_001		
GRLD304002_13_002	GRLD307003_13_002		
GRLD304002_13_003	GRLD307003_13_003		
GRLD304002_13_004	GRLD307003_13_004		
GRLD304002_13_006	GRLD307003_13_006		
GRLD304002_13_007	GRLD308001_13_001		
GRLD304002_13_008	GRLD308001_13_002		
GRLD304002_13_009	GRLD308001_13_003		
GRLD304002_13_011	GRLD308001_13_004		
GRLD304002_13_012	GRLD308001_13_006		
GRLD304002_13_013	GRLD308001_13_007		
GRLD307001_13_002	GRLD308001_13_008		
GRLD307001_13_003	GRLD308001_13_009		
GRLD307001_13_004	GRLD308001_13_010		
GRLD307001_13_006	GRLD308001_13_012		
GRLD307001_13_008	GRLD308001_13_013		
GRLD307001_13_009	GRLD308001_13_014		
GRLD307001_13_010	GRLD308001_13_015		
GRLD307001_13_011	GRLD308001_13_016		
GRLD307001_13_012	GRLD308001_13_017		
GRLD307001_13_013	GRLD308001_13_019		
GRLD307001_13_014	GRLD308001_13_020		
GRLD307001_13_015	GRLD308001_13_021		
GRLD307001_13_018	GRLD308001_13_022		
GRLD307001_13_019	GRLD308001_13_023		
GRLD307001_13_020	GRLD308001_13_024		
GRLD307001_13_021	GRLD308001_13_025		
GRLD307001_13_022	GRLD308001_13_027		
GRLD307001_13_025	GRLD308001_13_028		
GRLD307001_13_027	GRLD308001_13_029		

Table 6-7: plot numbers of validation plots, Douglas-fir.

GRLD020001_11_001	GRLD307001_13_001
GRLD020001_11_007	GRLD307001_13_005
GRLD020001_11_013	GRLD307001_13_016
GRLD020001_11_014	GRLD307001_13_017
GRLD020001_11_020	GRLD307001_13_023
GRLD021001_11_002	GRLD307001_13_034
GRLD021001_11_003	GRLD307001_13_036
GRLD021001_11_008	GRLD307001_13_038
GRLD026001_11_002	GRLD307001_13_040
GRLD026001_11_008	GRLD307001_13_043
GRLD026001_11_011	GRLD307003_13_005
GRLD026001_11_013	GRLD308001_13_005
GRLD026001_11_022	GRLD307001_13_024
GRLD033001_15_010	GRLD308001_13_011
GRLD033001_15_012	GRLD307001_13_026
GRLD043001_11_002	GRLD308001_13_018
GRLD103001_13_011	GRLD308001_13_034
GRLD103001_15_012	GRLD308001_13_037
GRLD105001_15_017	GRLD209001_12_005
GRLD105002_15_028	
GRLD110001_13_004	
GRLD110001_13_005	
GRLD110001_13_007	
GRLD110001_13_008	
GRLD111002_13_001	
GRLD111002_13_005	
GRLD111002_13_009	
GRLD111002_13_011	
GRLD111002_13_012	
GRLD207002_13_003	
GRLD207002_13_010	
GRLD207002_13_012	
GRLD208001_13_008	
GRLD208001_13_025	
GRLD209001_12_020	
GRLD209001_12_021	
GRLD209001_12_027	
GRLD209002_13_001	
GRLD209002_13_011	
GRLD304002_13_005	
GRLD304002_13_010	

6.4 Climate data

6.4.1 Analysis of wind direction

This analysis is provided to support contentions made in Chapter 4: *Discussion* regarding the effects of aspect and lee slopes.

The wind records for the Timaru aerodrome, the nearest weather station with wind records across the study period, were analysed for frequency by direction. The results are shown below in Figure 6-11, which shows all the recorded windspeeds, and Figure 6-12[7], which shows the top 2% of recorded windspeeds.

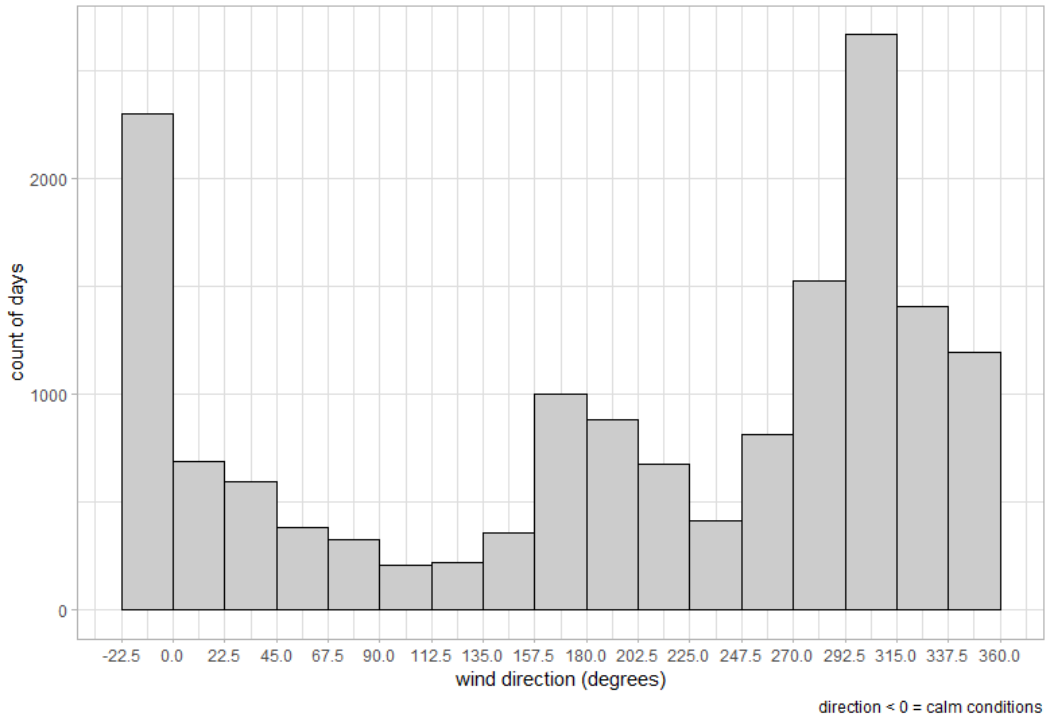

Figure 6-11: cumulative frequency of daily 9 am wind direction records for Timaru Aerodrome, 31/12/1971 to 31/12/2016.

[7] When reading this account of wind direction, it should be understood that groups of wind directions include the first direction given and move clockwise to the last direction given. For example, 'from north-east to south' means the directions north-east, east, south-east, and south, and 'from 180 degrees to 315 degrees' includes all directions in the set bounded by the two figures given.

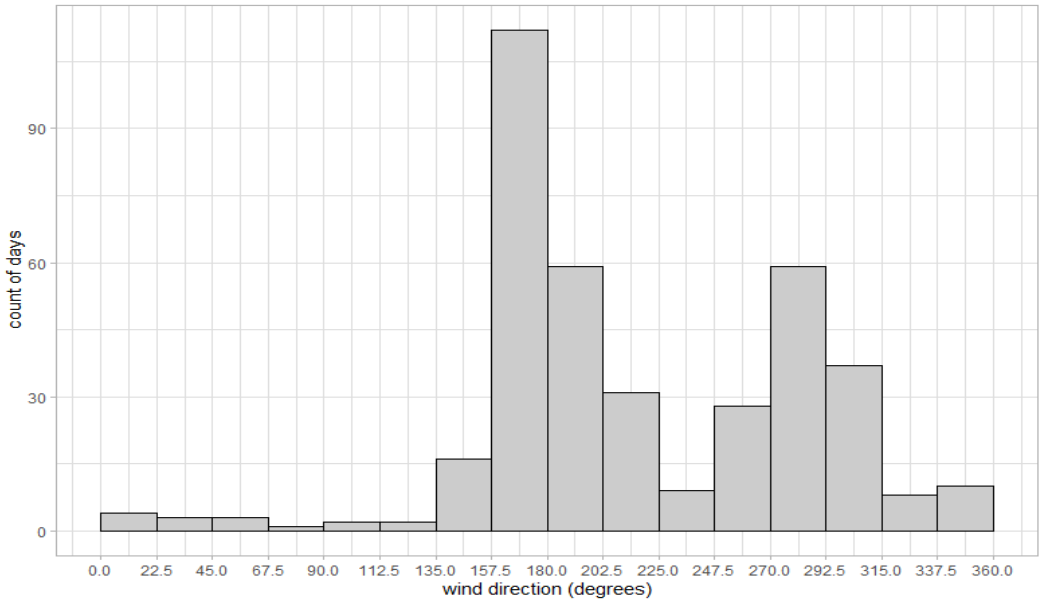

Figure 6-12: cumulative frequency of top 2% (29.7 km/hr and above) of daily 9 am wind direction records for Timaru Aerodrome, 31/12/1971 to 31/12/2016, by wind direction.

Clearly, winds are not evenly distributed. For all wind records there are frequency peaks at 157.5 – 180 degrees (south-south-east to south), and 292.5 - 315 degrees (west-north-west to north-west), with winds from the north-west quadrant dominant. For the top two percent of windspeeds, there is again a frequency peak at 157.5 – 180 degrees (south-south-east to south), and there is another at 270 – 292.5 degrees (west-north-west to north-west), but in this case winds from the south quadrant are dominant.

Plotting the top two percent of wind speeds for frequency by month shows that the peak months for strong winds are October and November, as shown in Figure 6-13 below.

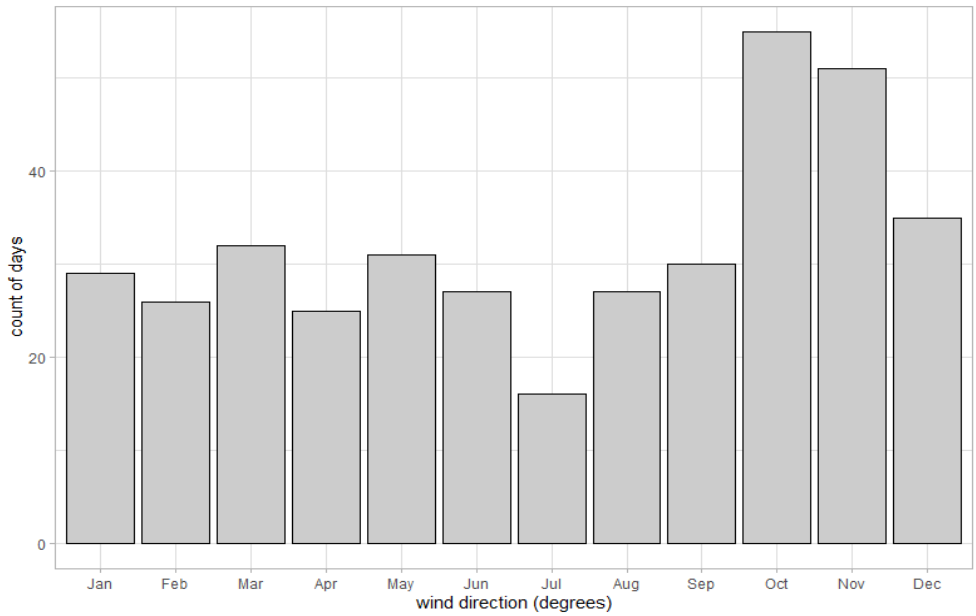

Figure 6-13: cumulative frequency of top 2% (29.7 km/hr and above) of daily 9 am wind direction records for Timaru Aerodrome, 31/12/1971 to 31/12/2016, by month of year.

6.4.2 Virtual Climate Station Network data

Virtual Climate Station number 15231 was chosen as being reasonably representative for Geraldine Forest. It is located at 44.1250 S 171.0828 E. This is towards the lower end of the elevation range for Geraldine Forest, a point which has been considered when setting thresholds. All data for the site were extracted from the earliest available date, 31/12/1971, to 24/11/2018, the date on which the data were extracted.

Maximum and minimum daily temperature, precipitation over 24 hours, and air pressure at 9 am are available across this date range. Wind data, however, do not commence until 01/01/1997. Therefore, wind data were obtained for Timaru Aerodrome, and compiled as described below.

6.4.3 Timaru Aerodrome wind speed data

Three different data sources were compiled to obtain a full windspeed data set for Timaru Aerodrome. The first source is Timaru Aerodrome data from 31/01/1971[8] to 04/06/1990, sourced from NIWA's CliFlo database. These were (presumably) from a manual weather station. From 04/06/1990 to 1/01/2013, the data are for Timaru Aerodrome automatic weather station, sourced from NIWA's CliFlo database. From 1/01/2013 until 31/12/2016, data are from the same automatic weather station, but sourced instead from the MetService (the New Zealand Government metrological department), who control the data from that date.

Before 02/11/2009, average windspeed was, in most cases, calculated over a 10-minute interval ending at 9 am. The average was then calculated hourly until 01/01/2013, then switched back to the 10-minute interval. Prior to 2/11/09 windspeeds were recorded in whole knots. From 3/11/2009 to 31/12/2012, windspeeds were recorded in km/hr to one decimal place. From 01/01/2013 onwards, windspeeds were recorded in whole km/hr. All data used here have been converted to km/hr, but the difference recording practices give artificial consistency beyond the decimal place. Some dates during the period of this study have no windspeed data. These are shown in Table 6-8, below.

Table 6-8: dates for which there are no Timaru Aerodrome windspeed data.

25/10/1986	24/06/1990	7/04/1991	20/04/1991	3/05/1991	16/05/1991	28/05/1996
24/10/1989	25/06/1990	8/04/1991	21/04/1991	4/05/1991	17/05/1991	23/10/1996
25/01/1990	1/07/1990	9/04/1991	22/04/1991	5/05/1991	18/05/1991	15/02/2004
16/02/1990	14/07/1990	10/04/1991	23/04/1991	6/05/1991	19/05/1991	16/02/2004
17/02/1990	10/10/1990	11/04/1991	24/04/1991	7/05/1991	20/05/1991	17/02/2004
2/04/1990	2/11/1990	12/04/1991	25/04/1991	8/05/1991	21/05/1991	13/11/2006
16/04/1990	15/11/1990	13/04/1991	26/04/1991	9/05/1991	22/05/1991	14/03/2007
13/05/1990	1/04/1991	14/04/1991	27/04/1991	10/05/1991	18/07/1991	20/06/2008
23/05/1990	2/04/1991	15/04/1991	28/04/1991	11/05/1991	19/07/1991	8/07/2008
28/05/1990	3/04/1991	16/04/1991	29/04/1991	12/05/1991	20/07/1991	9/07/2008
1/06/1990	4/04/1991	17/04/1991	30/04/1991	13/05/1991	21/07/1991	15/08/2008
2/06/1990	5/04/1991	18/04/1991	1/05/1991	14/05/1991	20/10/1991	24/11/2009
3/06/1990	6/04/1991	19/04/1991	2/05/1991	15/05/1991	7/02/1993	18/06/2013

6.4.4 Cropping the bottom end of the data set for weather variable calculation

Climate data from the VCSN are available from 31/12/1971, and the windspeed data from Timaru Aerodrome are available from 01/01/1970; but the planted date of the trees included in this study ranges back to 01/07/1962. For calculation of variables that express how much adverse weather a stand has experienced, the variable must apply to a consistent amount of the stand's life. Ideally this would be planting until measurement, but using planting date in this manner removes 75 plots from the input data, which is an unacceptably high loss, especially as these are all the plots for five stands, and thus removing them removes the coverage of the data set in some geographic areas. Instead, variables expressing adverse weather events are calculated from age 5. This reduces the stands lost to modelling to 25, in two stands. This choice of age 5 follows Somerville (1995), who considered that wind damage to stands under 5 years old would be largely in the form of leaning stems, not breakage or windthrow.

6.4.5 More extensive weather data for Virtual Climate Station number 15231

Table 6-9, below, shows a the minimum, mean, maximum and standard deviation for a selection of important climate metrics averaged over the period 1 January 1997 (start of full VCSN data at this station) to 31 December 2016 (end of the study period'. These figures constitute a fair representation of the 'average' climate at Geraldine Forest.

Table 6-9: Weather data for 1 January 1997 to 31 December 2016, for Virtual Climate Station number 15231.
Figures are daily unless otherwise stated.

	maximum temperature (°C)				**minimum temperature (°C)**				**24 hr mean windspeed (km/hr)**				**accumulated daily precipitation (mm)**				**accumulated monthly precipitation (mm)**
month	mean	max.	min.	s.d.	mean	max.	min.	s.d.	mean	max.	min.	s.d.	mean	max.	min.	s.d.	mean
January	21.4	32.5	11.1	4.4	9.7	16.8	1.3	2.9	2.8	9	0.9	6.5	2.4	66.8	0	6.5	74.6
February	21.2	36.3	11.8	4.2	9.7	16.9	1.4	2.9	2.6	6.5	0.6	6.9	2.2	82.0	0	6.9	61.6
March	19.6	33.3	10.9	4.0	7.7	16.8	0.2	3.0	2.5	7.1	0.5	5.8	1.7	64.6	0	5.8	53.7
April	16.3	24.8	7.8	3.5	4.8	11.6	-2.9	3.0	2.1	6.5	0.0	6.8	2.3	84.4	0	6.8	68.7
May	13.4	25.4	5.7	3.6	2.7	11.1	-5.6	3.1	2.2	6.9	0.1	6.8	1.9	87.8	0	6.8	60.0
June	10.6	19.0	1.8	3.1	-0.4	6.7	-7.9	2.7	2.2	6.5	0.0	6.0	1.7	58.7	0	6.0	50.8
July	10.1	19.8	1.0	3.0	-0.9	7.9	-7.6	2.7	2.2	7.1	0.0	6.2	1.6	71.5	0	6.2	50.6
August	11.4	20.5	2.3	3.3	0.5	8.6	-6.2	2.7	2.3	7.4	0.0	8.3	2.1	145.0	0	8.3	65.1
September	14.4	25.8	5.1	3.7	2.6	10.9	-4.5	2.9	2.7	8.0	0.3	4.0	not available - source data error				
October	16.2	27.0	6.1	3.9	4.3	14.4	-2.5	3.0	3.0	9.8	0.7	5.4	2.1	56.2	0	5.4	66.5
November	17.9	29.2	8.3	4.3	6.1	15.6	-1.5	3.0	3.0	7.6	0.8	5.6	2.1	42.3	0	5.6	64.2
December	19.9	33.6	8.7	4.1	8.6	16.4	0.6	2.9	2.9	8.2	0.9	6.5	2.3	63.6	0	6.5	69.9

6.5 Exploratory data analysis

Throughout these figures, PRAD refers to radiata pine, and PSMEN to Douglas-fir.

6.5.1 Variables pertaining to individual trees

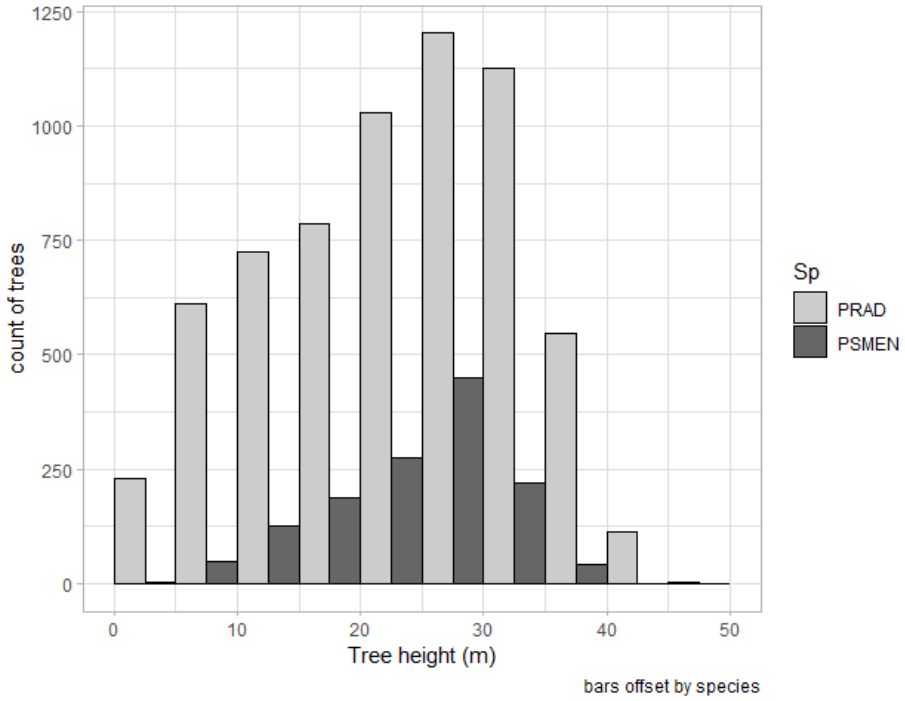

Figure 6-14: distribution of individual tree heights, by species.

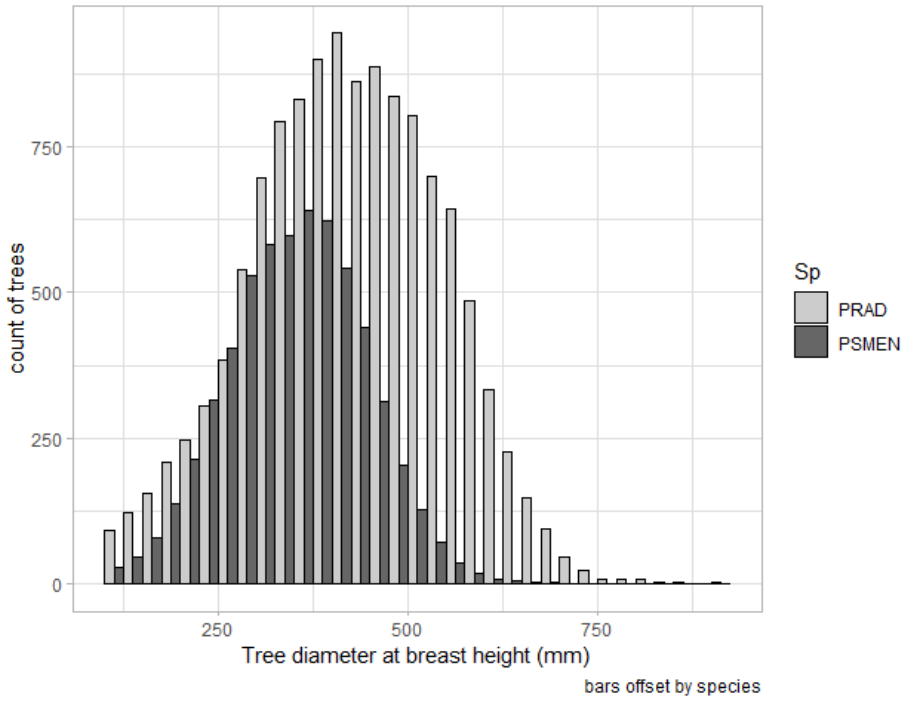

Figure 6-15: distribution of individual tree diameters at breast height, by species.

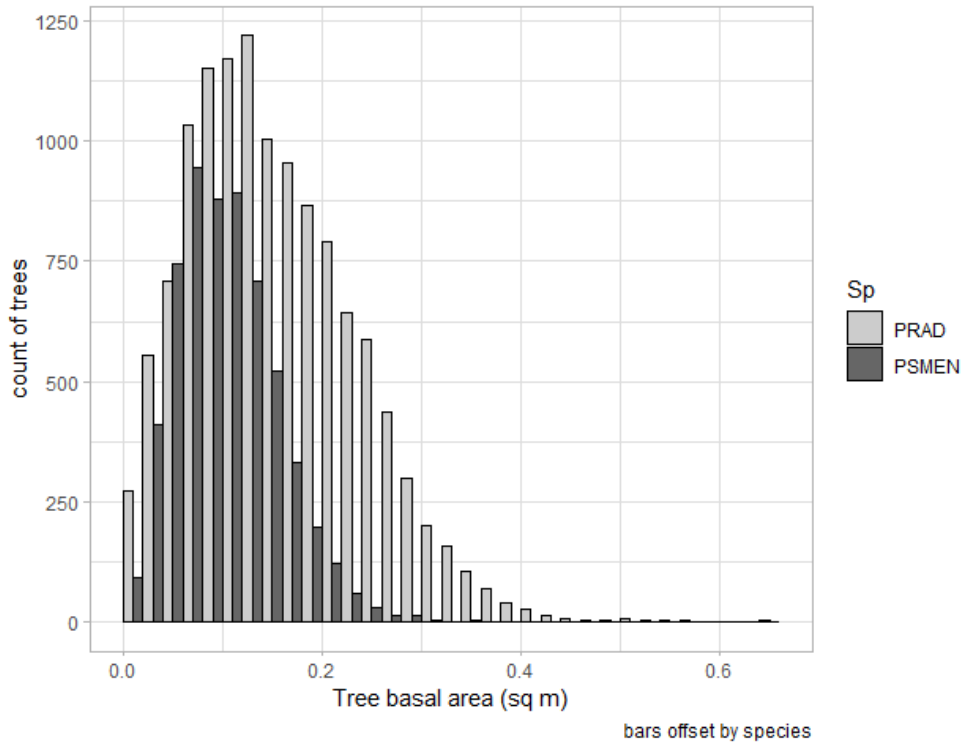

Figure 6-16: distribution of individual tree basal areas, by species.

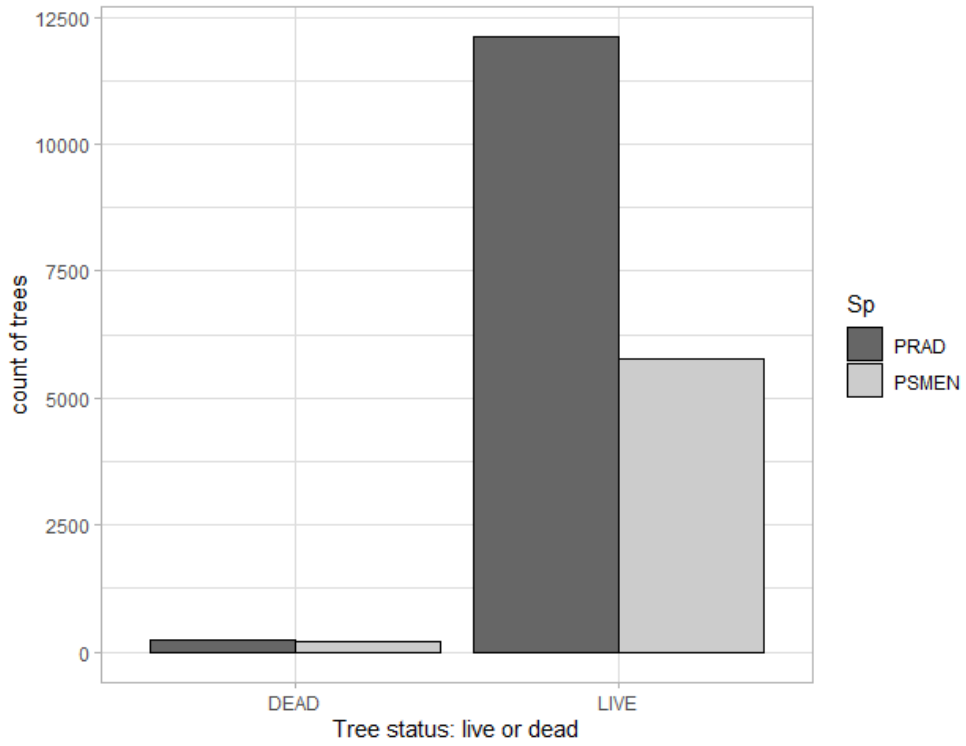

Figure 6-17: distribution of individual tree live/dead status, by species.

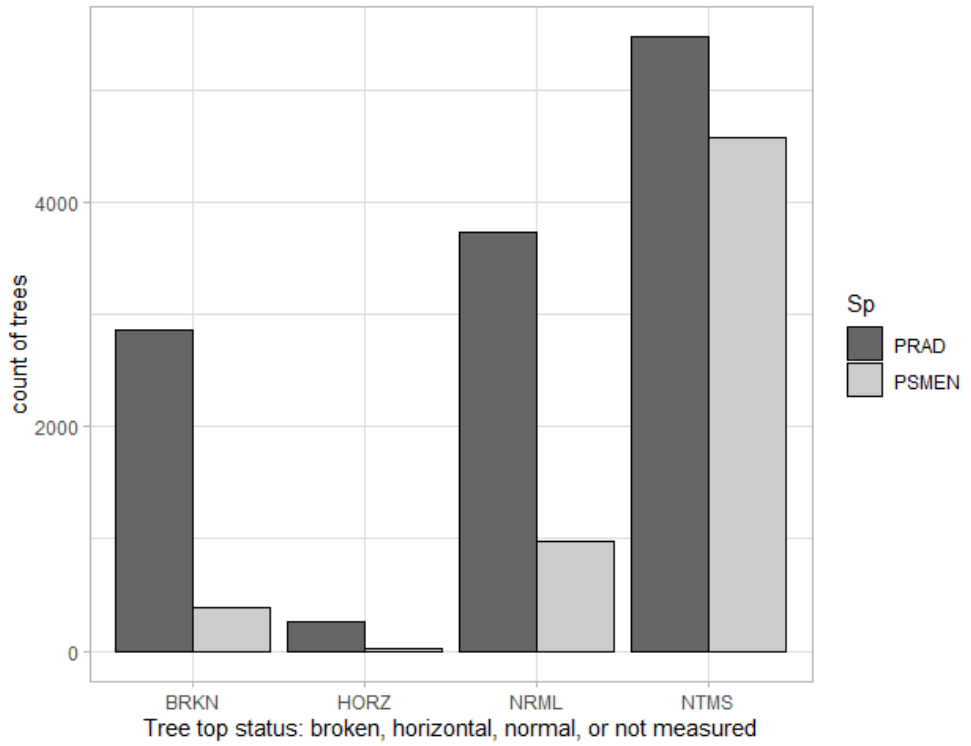

Figure 6-18: distribution of individual tree top status, broken/horizontal/normal/not measured.

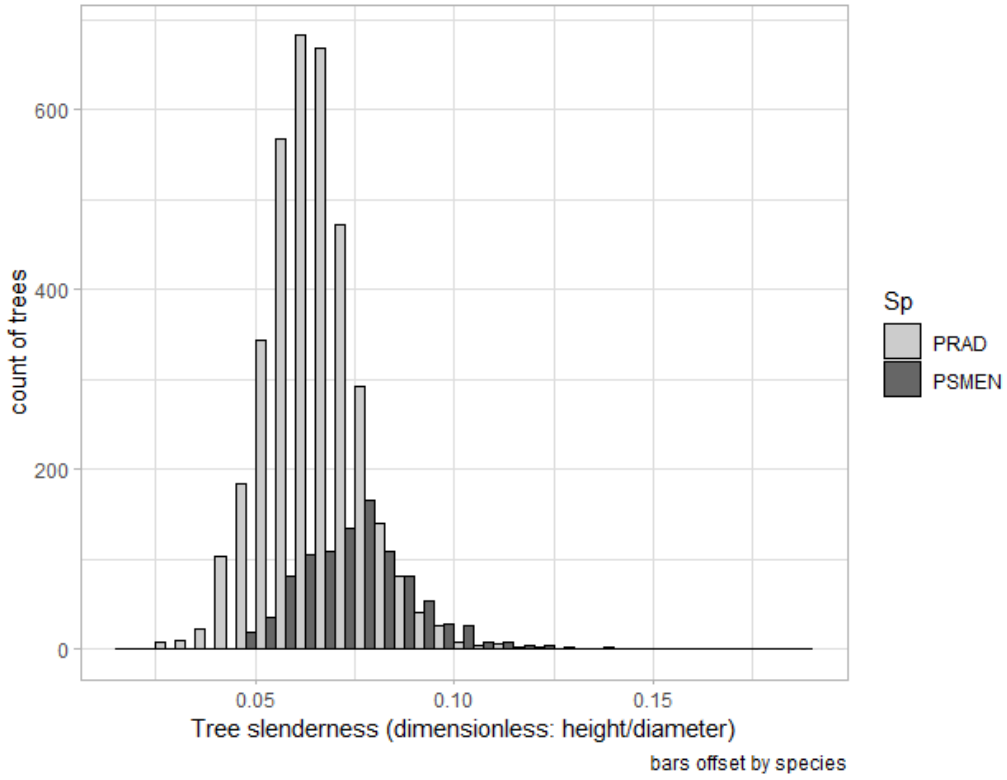

Figure 6-19: distribution of individual tree slenderness, by species.

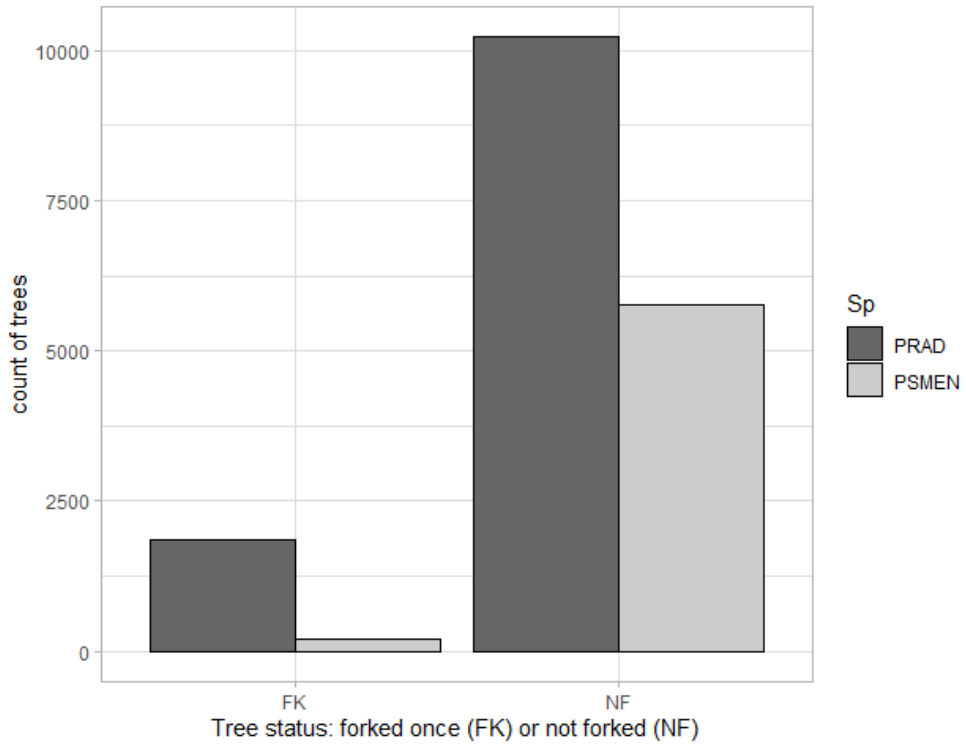

Figure 6-20: occurrence of first forks, by species.

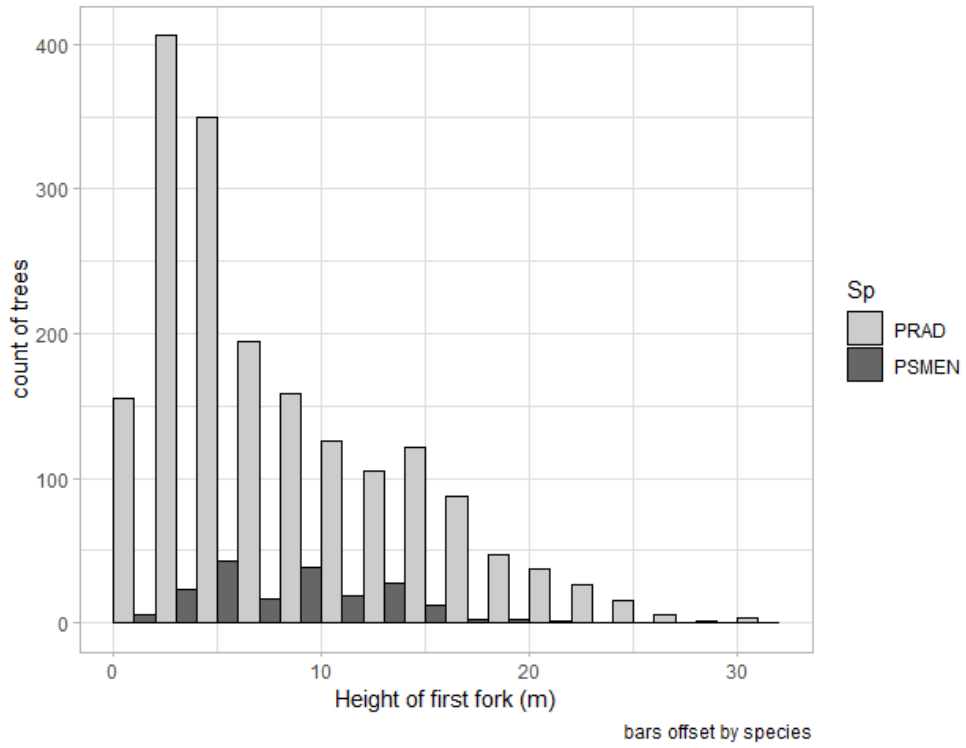

Figure 6-21: distribution of height of first fork, by species.

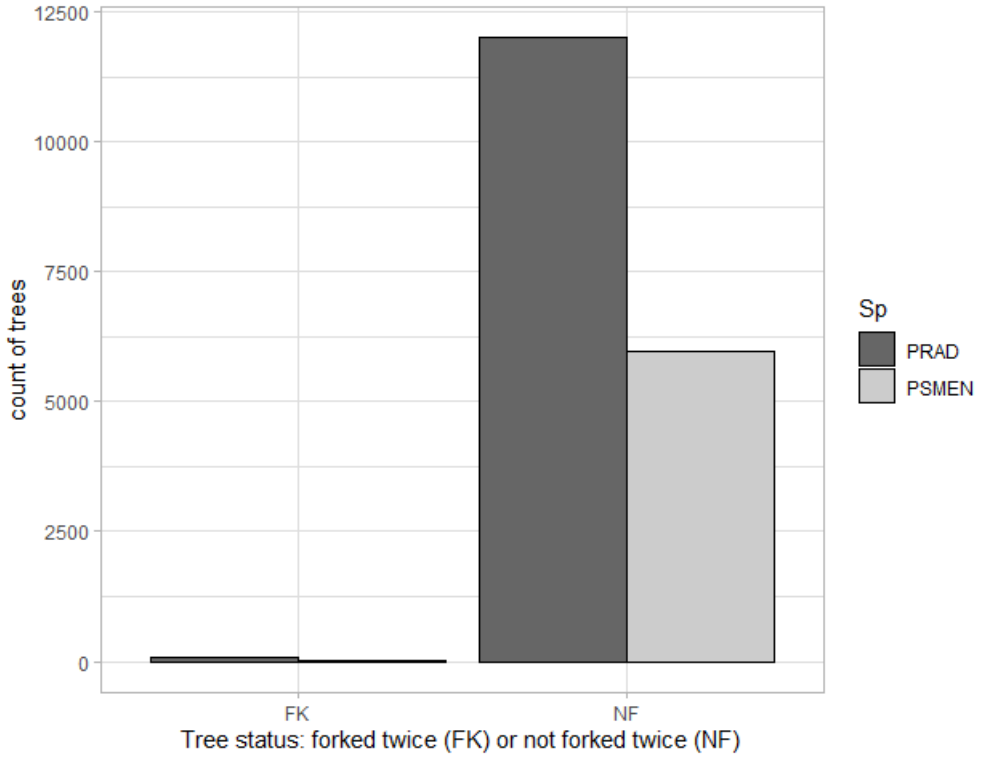

Figure 6-22: occurrence of second forks, by species.

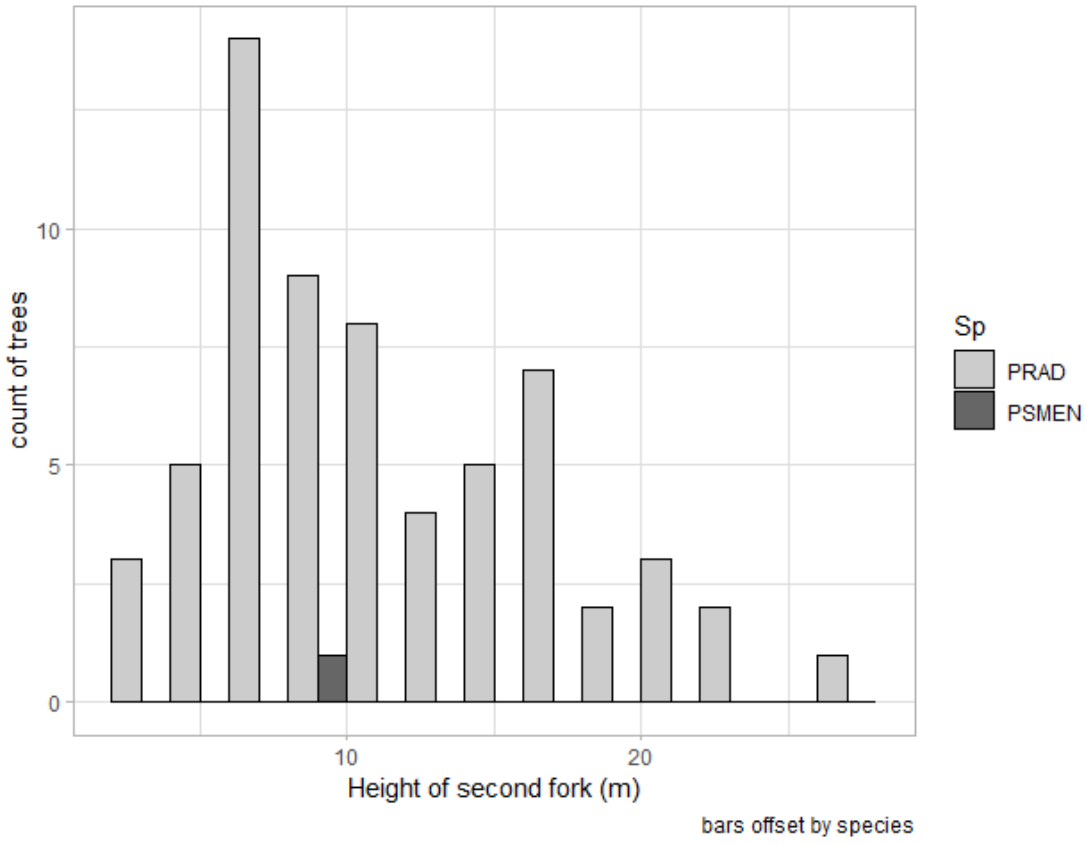

Figure 6-23: distribution of height of second fork, by species.

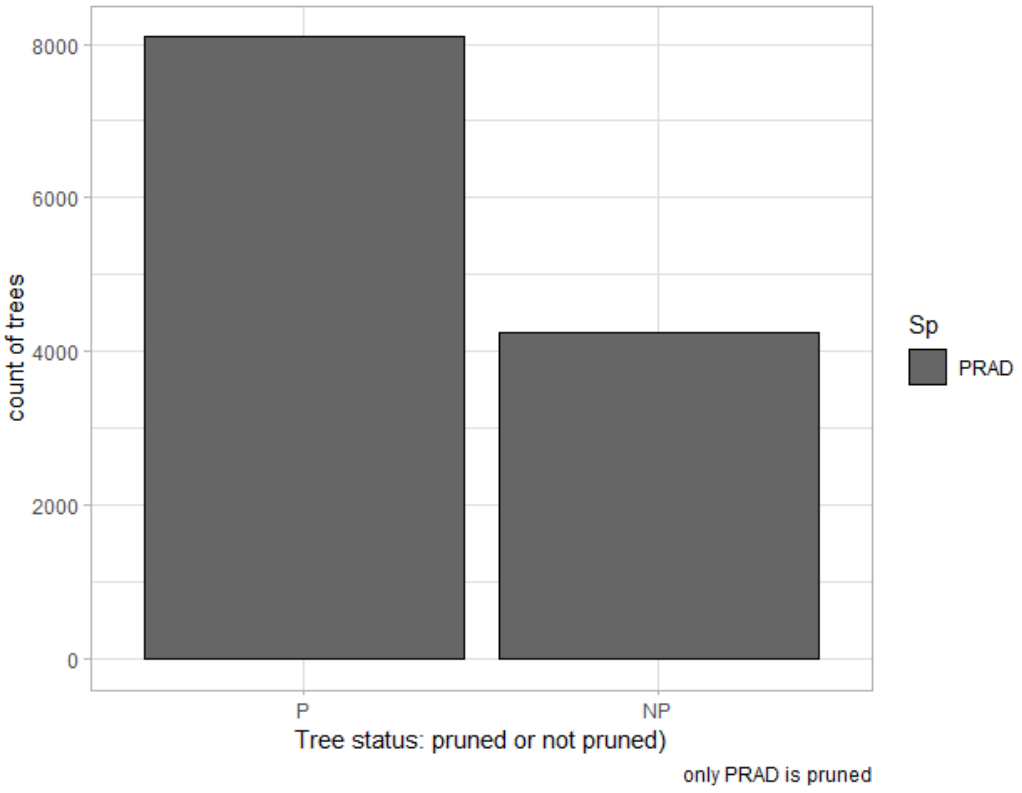

Figure 6-24: occurrence of pruning, PRAD only. PRAD = radiata pine.

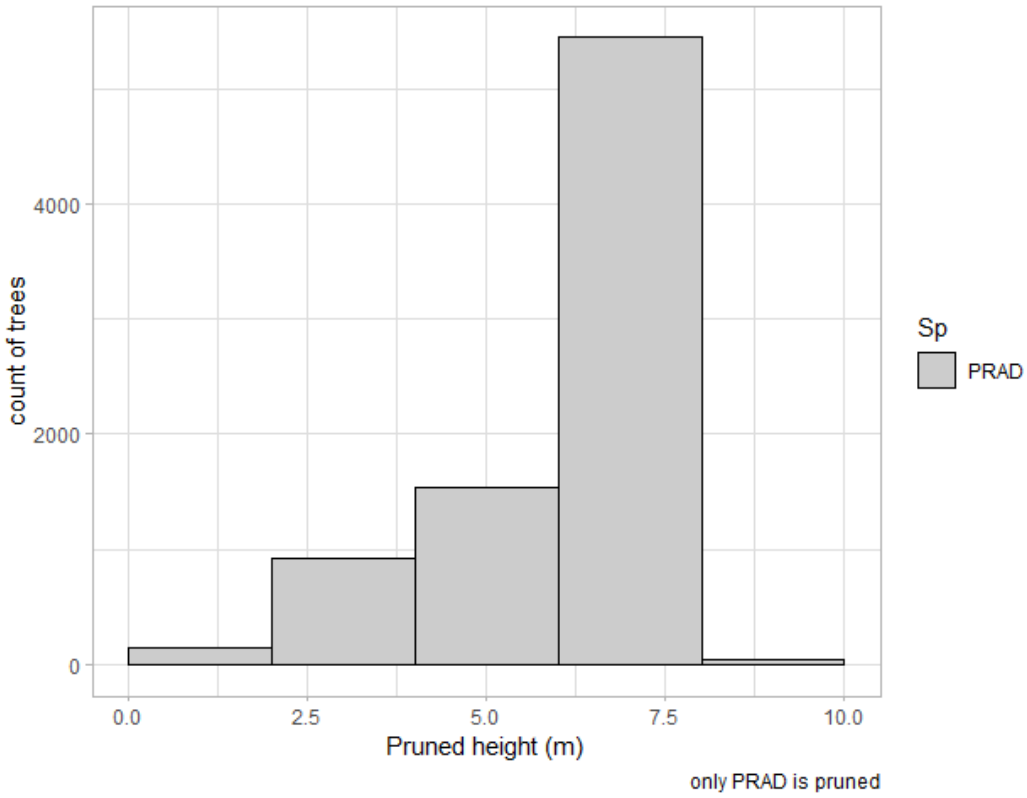

Figure 6-25: distribution of pruned heights, PRAD only. PRAD = radiata pine.

6.5.2 Variables calculated at the plot level

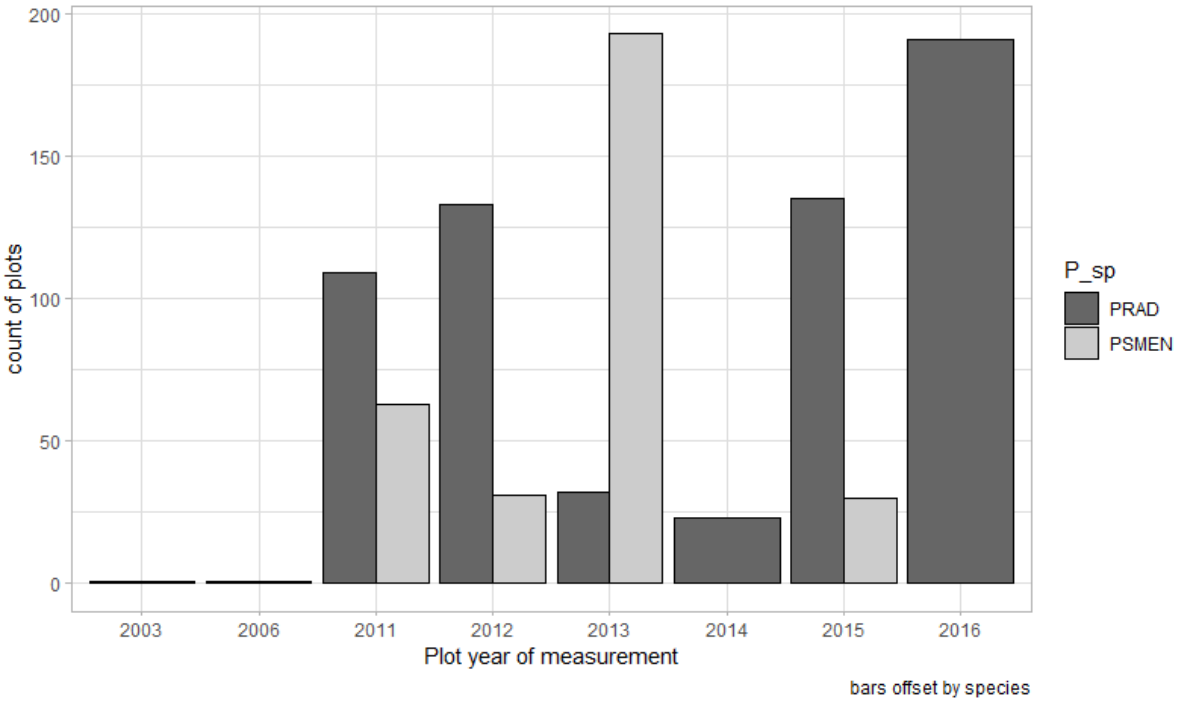

Figure 6-26: Frequency of plots by year of measurement (P_YOM).

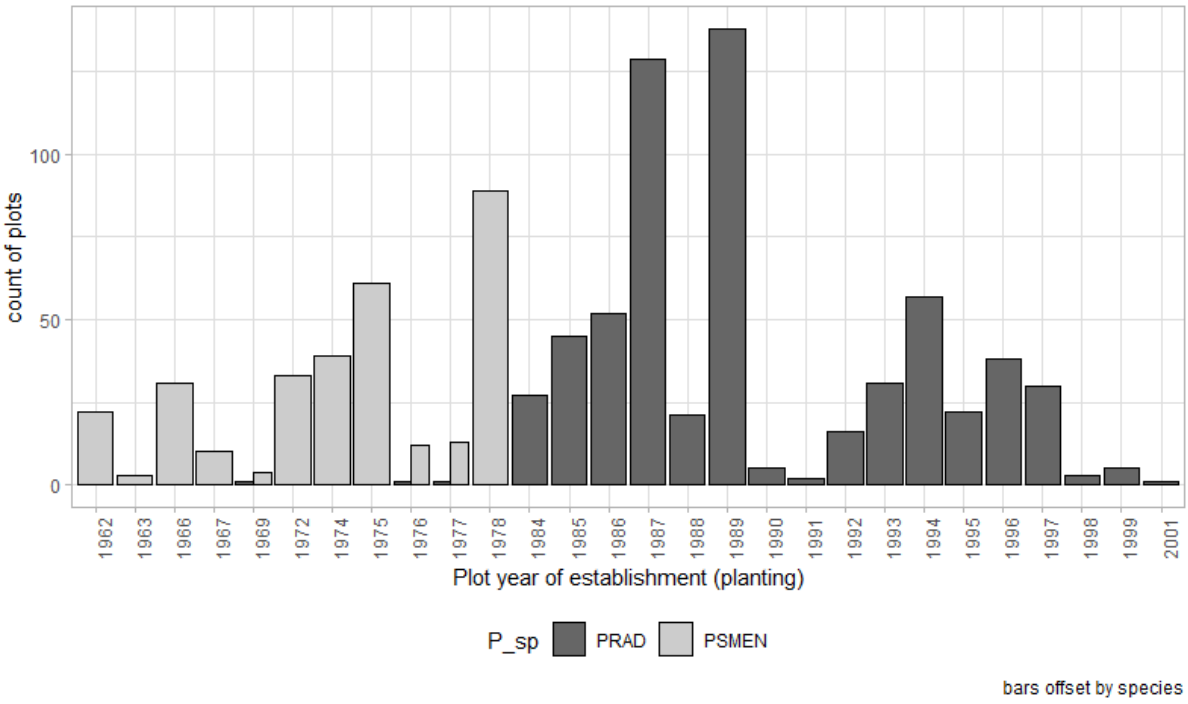

Figure 6-27: Frequency of plots by year of establishment (P_YOE).

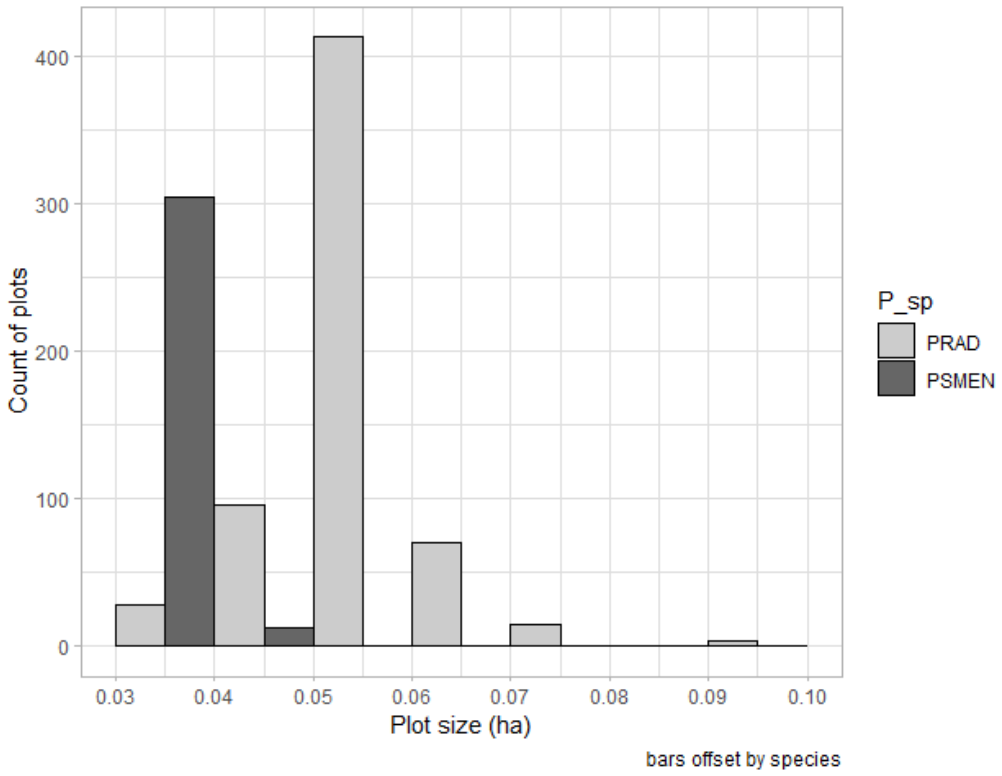

Figure 6-28: Distribution of plot sizes (P_size).

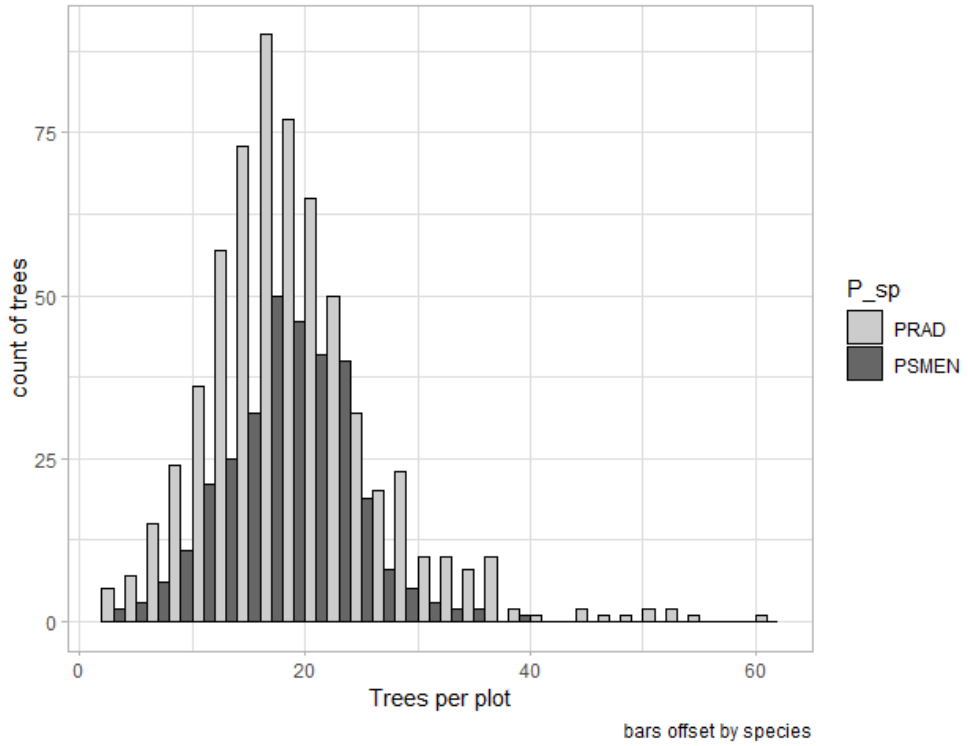

Figure 6-29: Distribution of trees per plot (P_count). Plot sizes differ: P_sph_equiv is generally more informative.

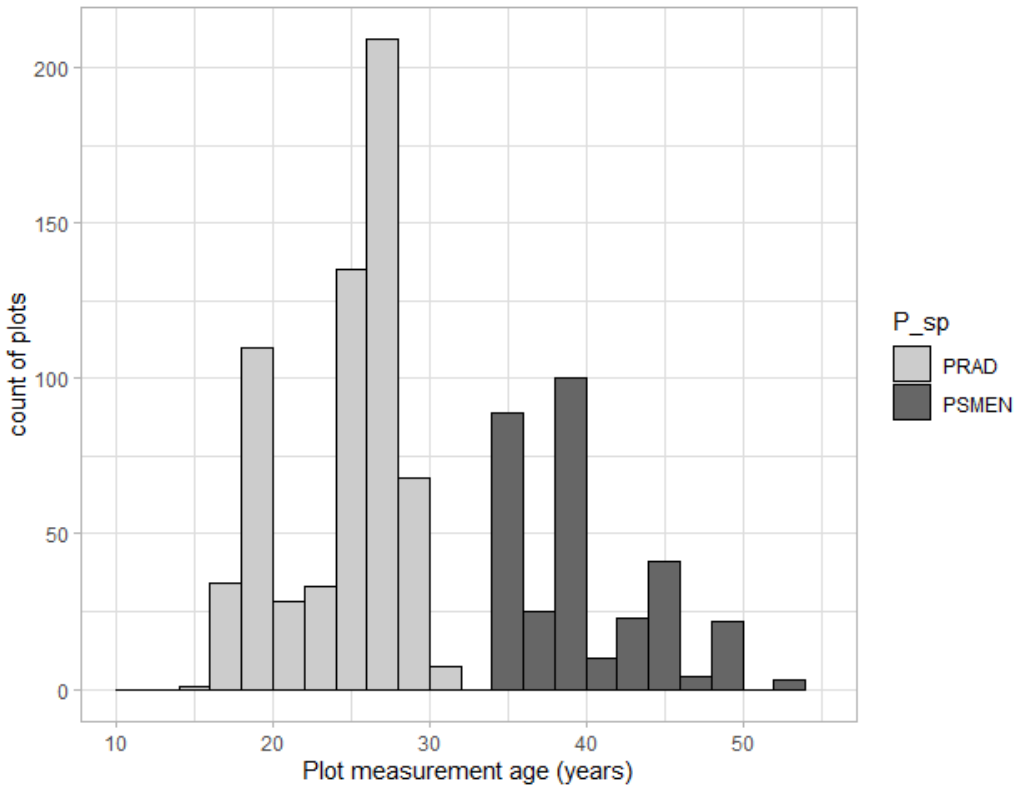

Figure 6-30: Distribution of plot measurement age (P_age_meas).

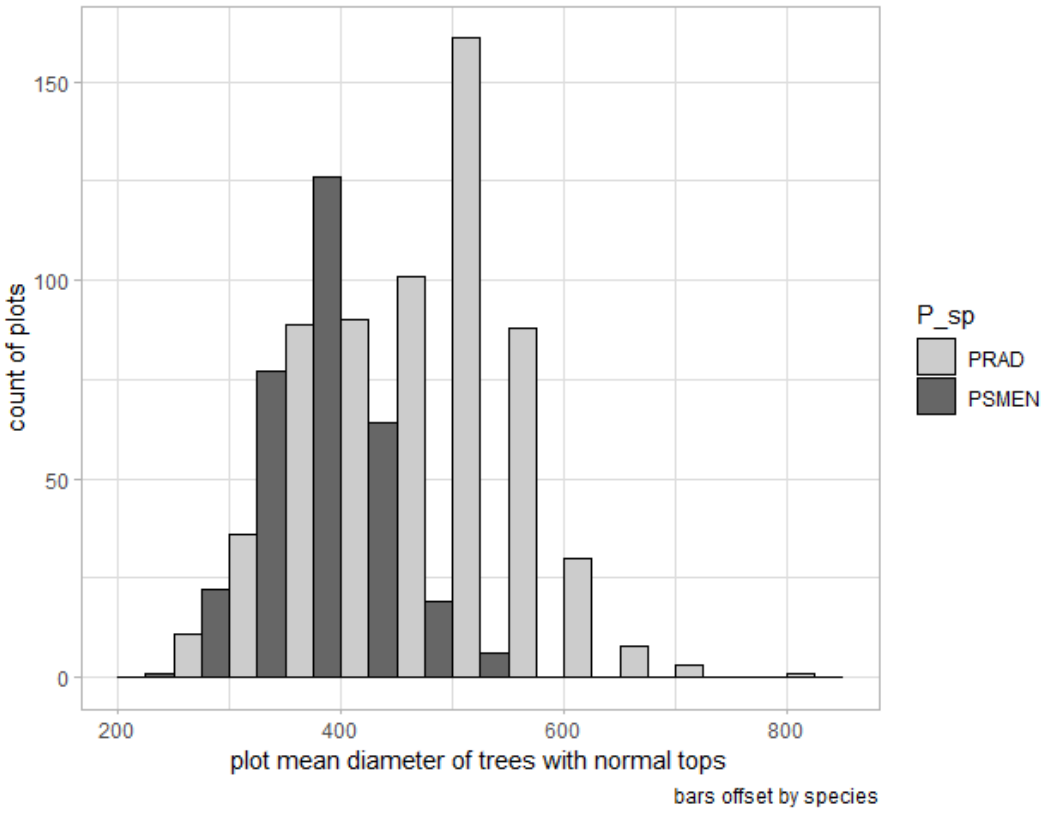

Figure 6-31: Distribution of plot mean diameter at breast height for trees with normal tops (P_dbh_mean_NRML).

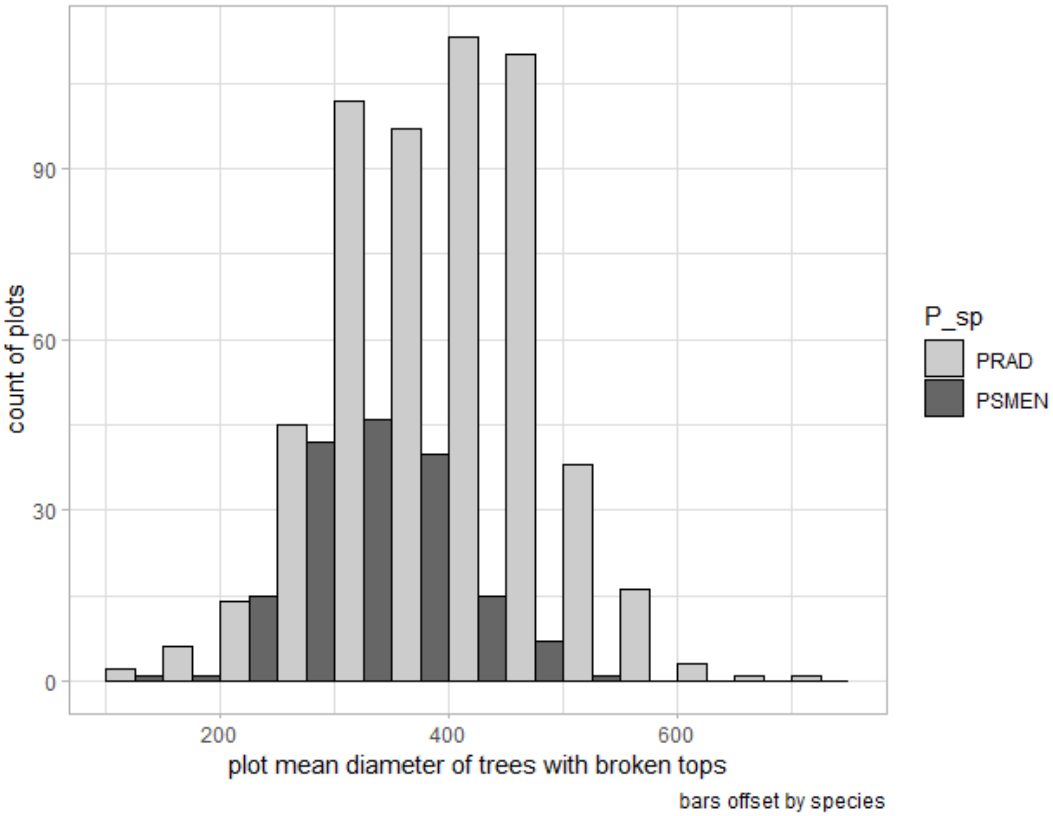

Figure 6-32: Distribution of plot mean diameter at breast height for trees with broken tops (P_dbh_mean_BRKN).

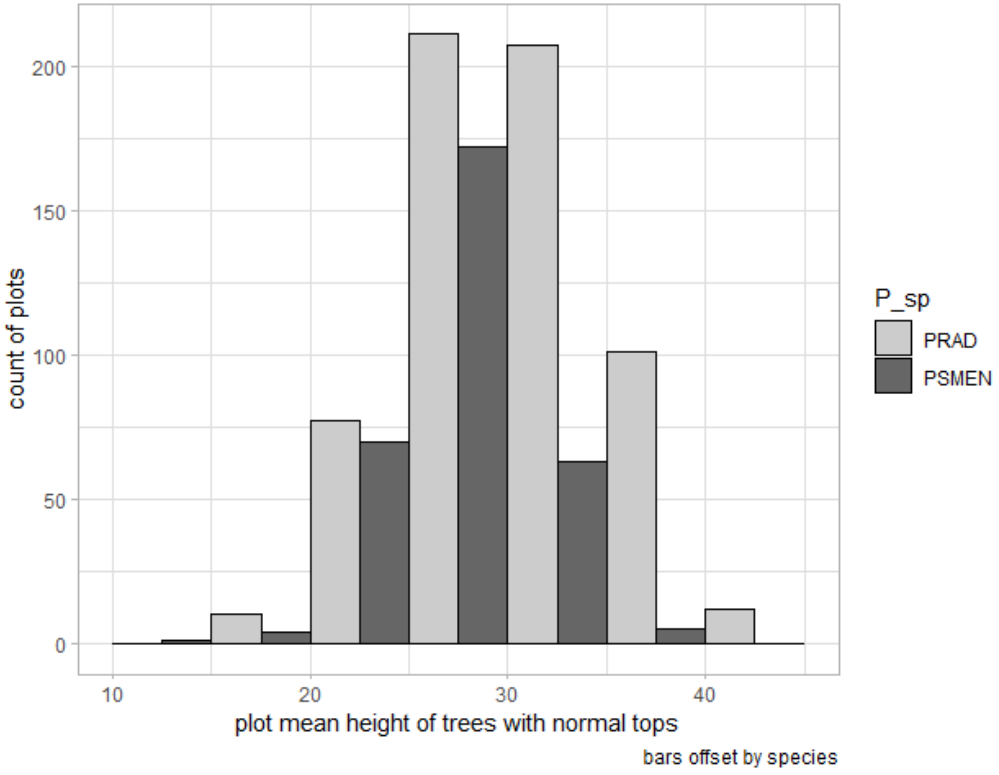

Figure 6-33: Distribution of tree heights for trees with normal tops (P_tree_ht_mean_NRML).

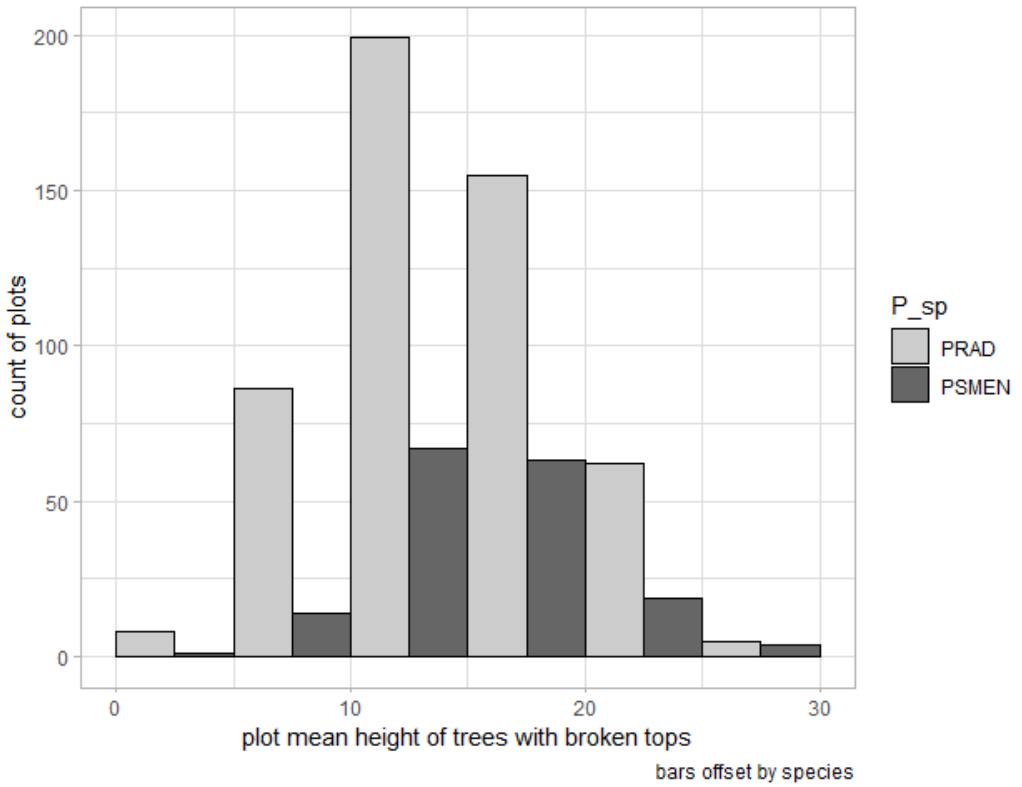

Figure 6-34: Distribution of tree heights for trees with broken tops.

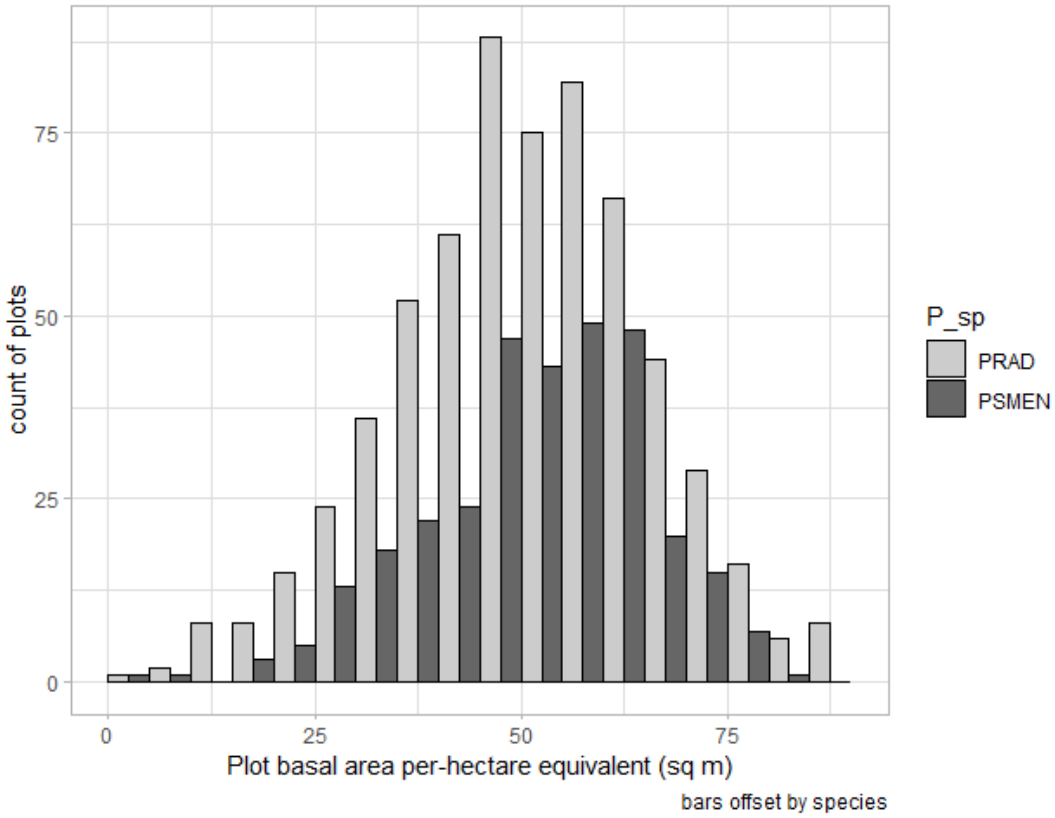

Figure 6-35: Distribution of plot basal area per-hectare equivalent (P_BA_ha_equiv).

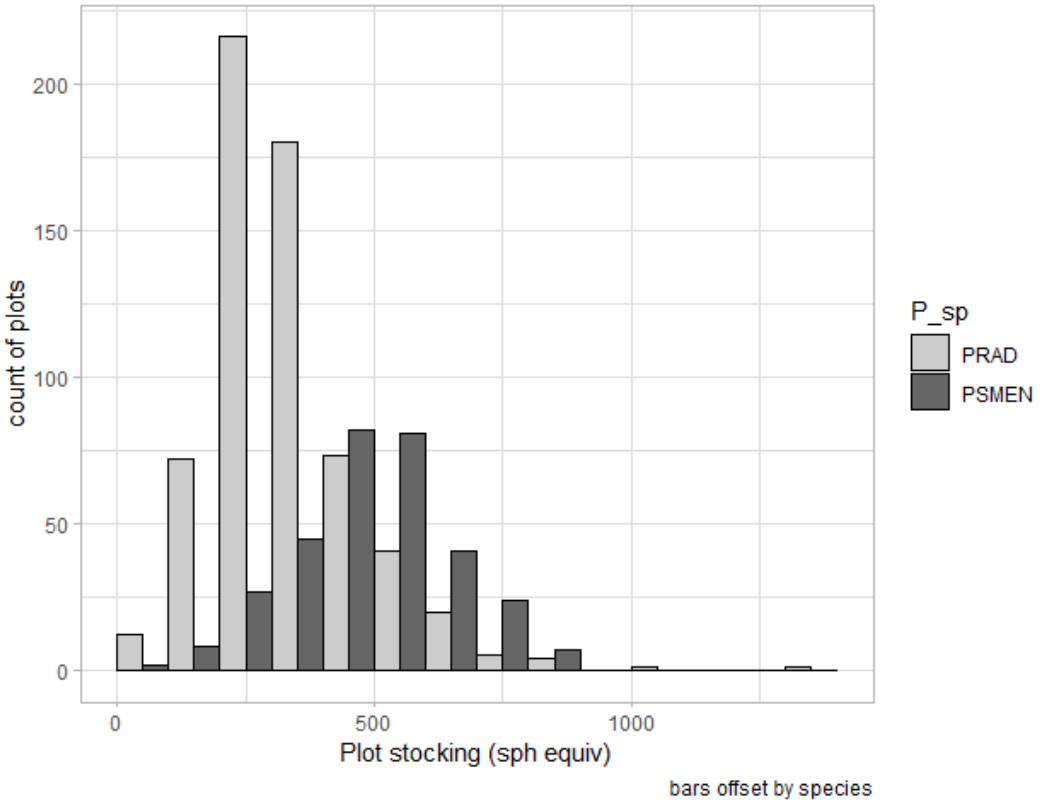

Figure 6-36: Distribution of plot stocking per-hectare equivalent (P_sph_equiv).

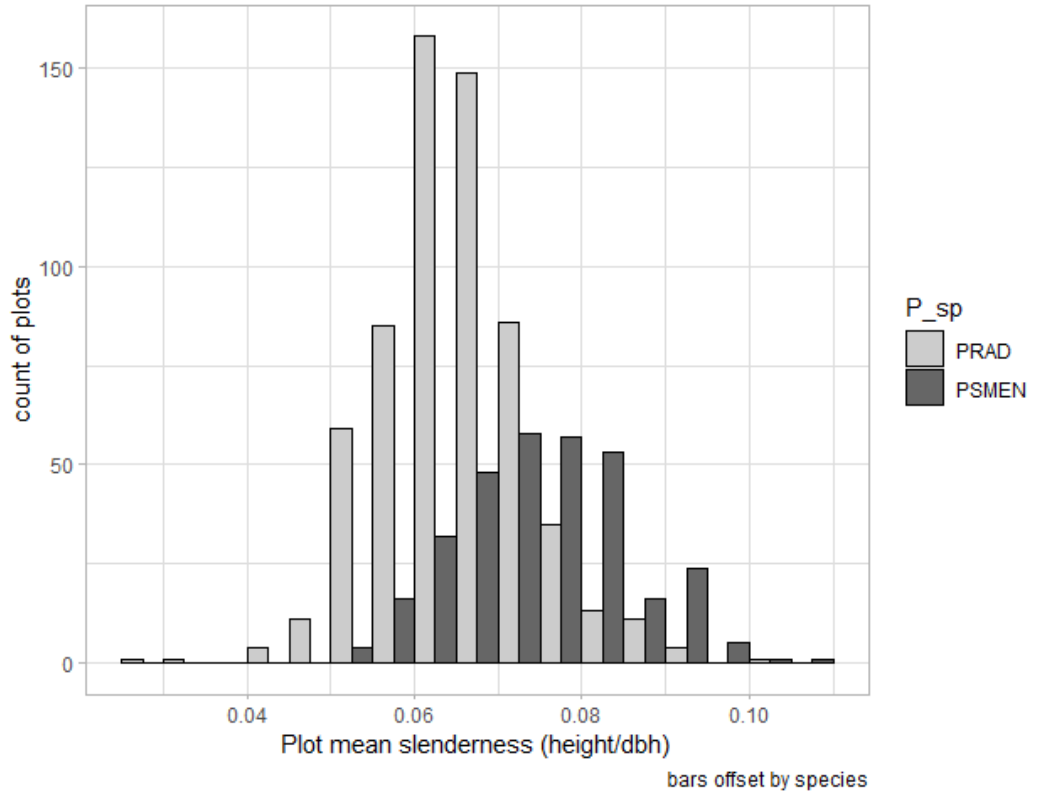

Figure 6-37: Distribution of plot mean slenderness (P_slend_mean).

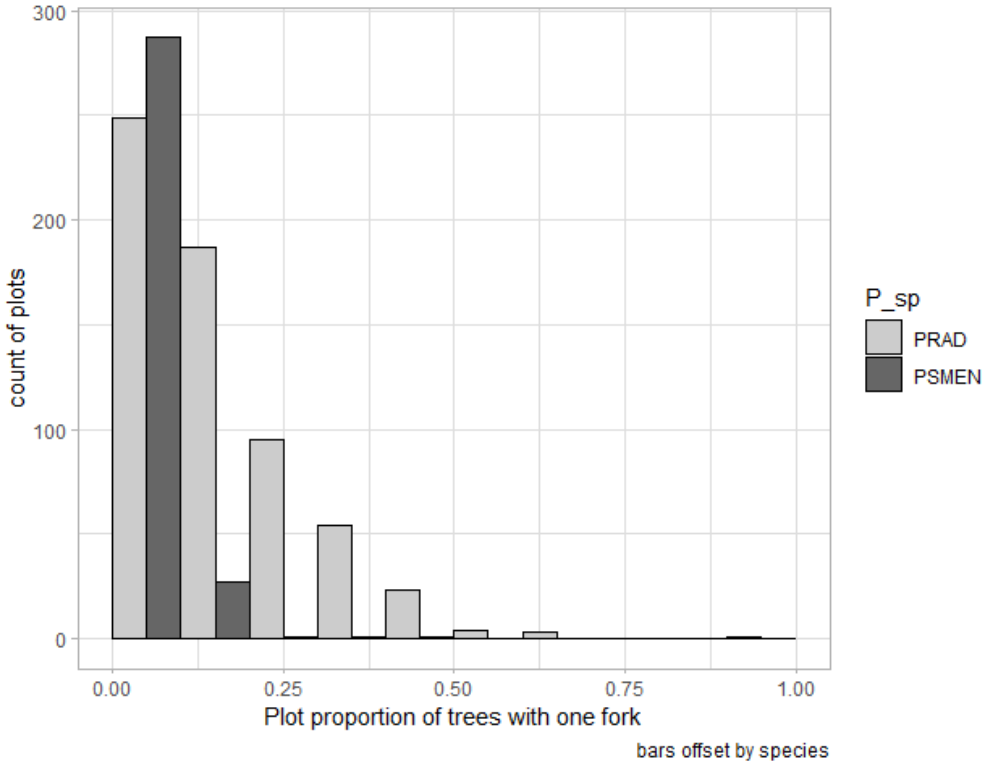

Figure 6-38: Distribution of proportion of trees per plot that have one fork.

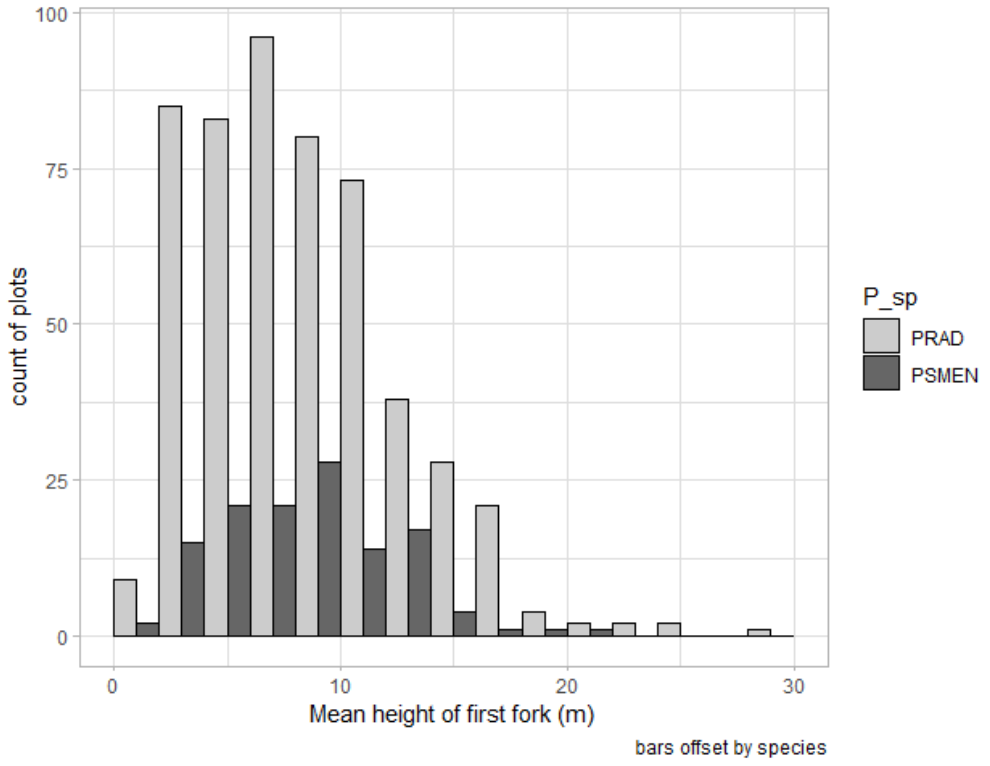

Figure 6-39: Distribution of per-plot mean height of the first fork.

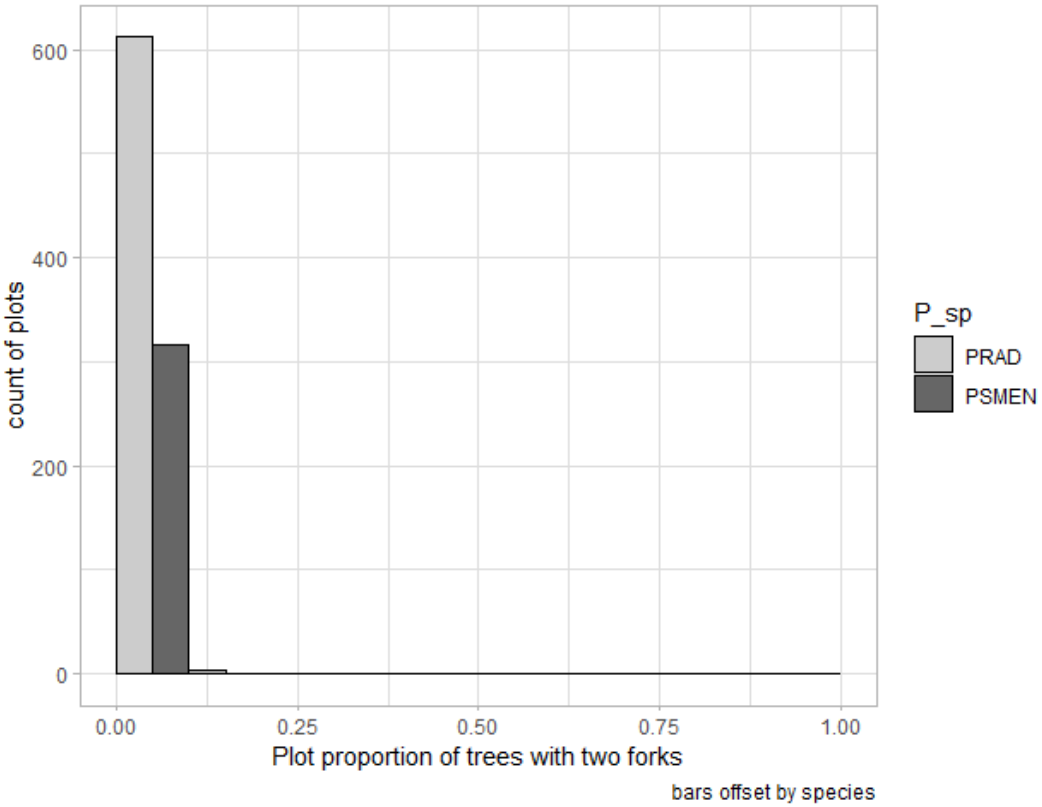

Figure 6-40: Distribution of proportion of trees per plot that have two forks.

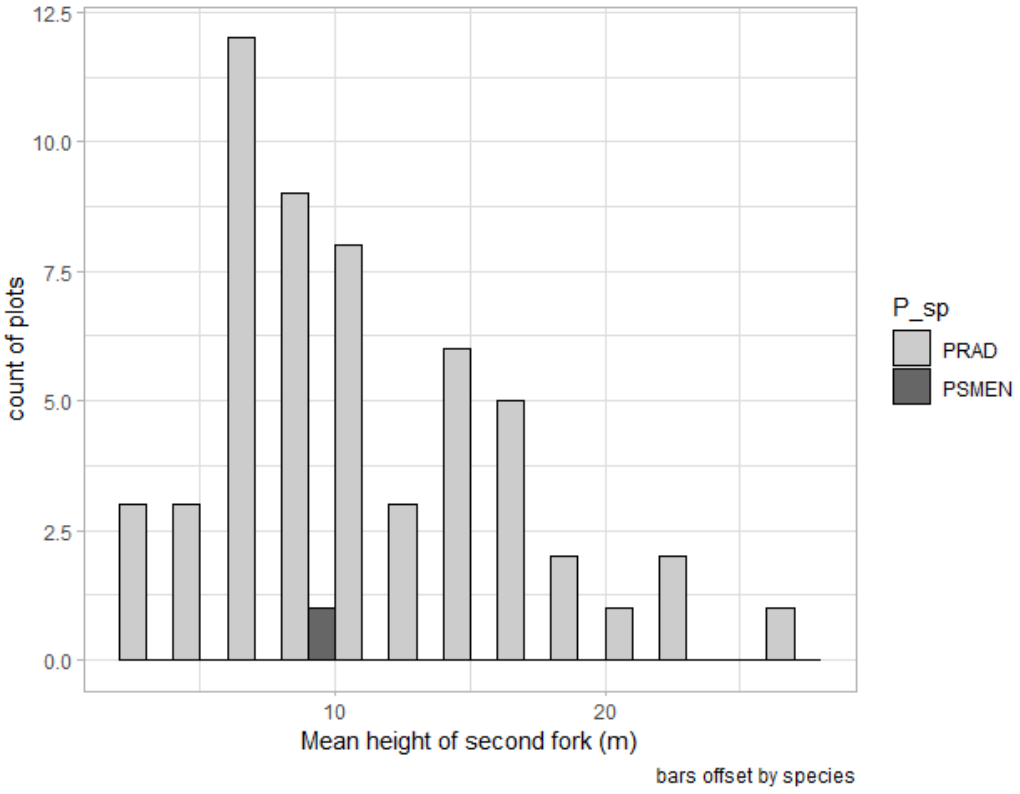

Figure 6-41: Distribution of per-plot mean height of the second fork.

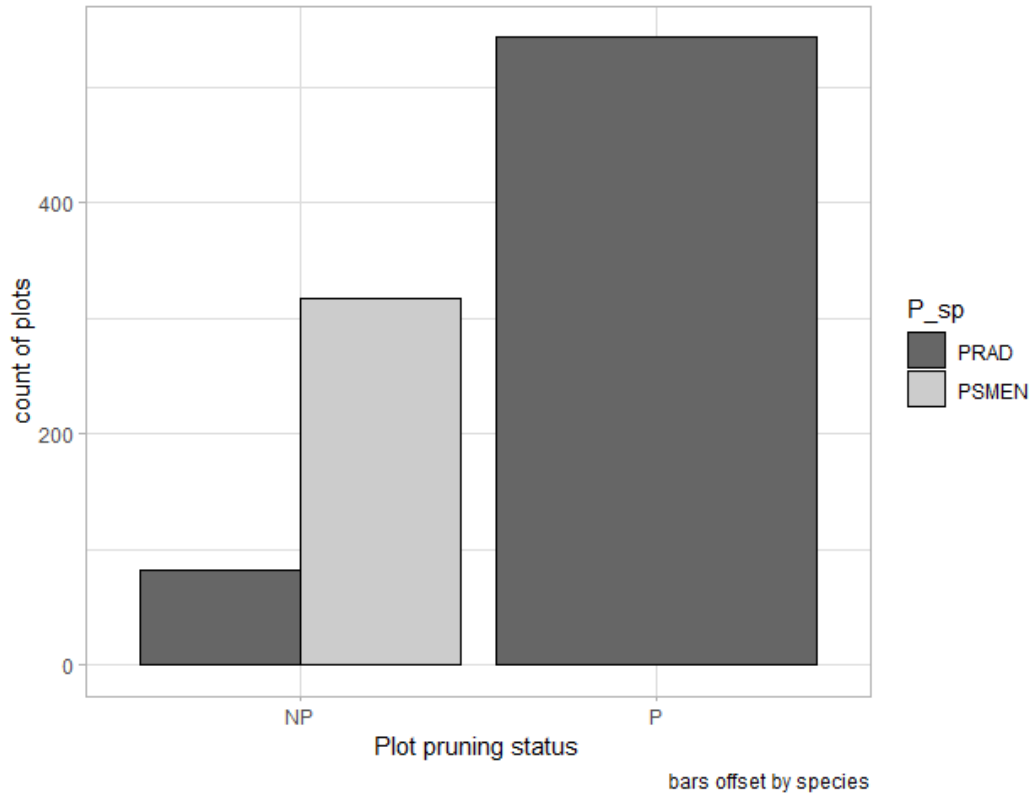

Figure 6-42: Plot pruned/unpruned status by species.

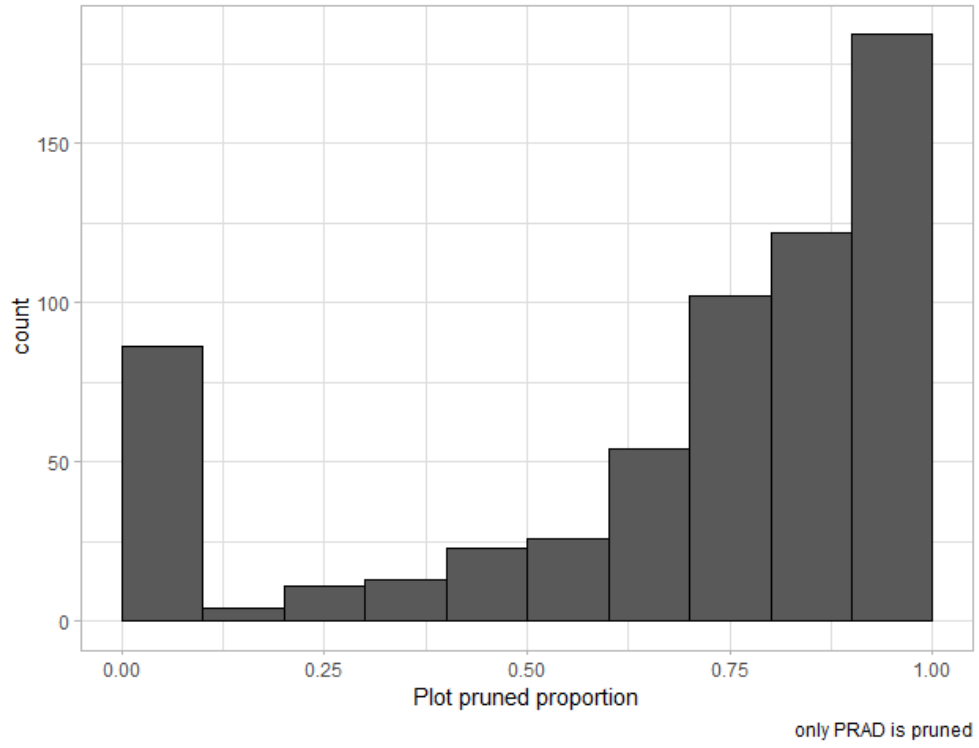

Figure 6-43: Distribution of proportion of trees pruned per plot. PRAD = radiata pine.

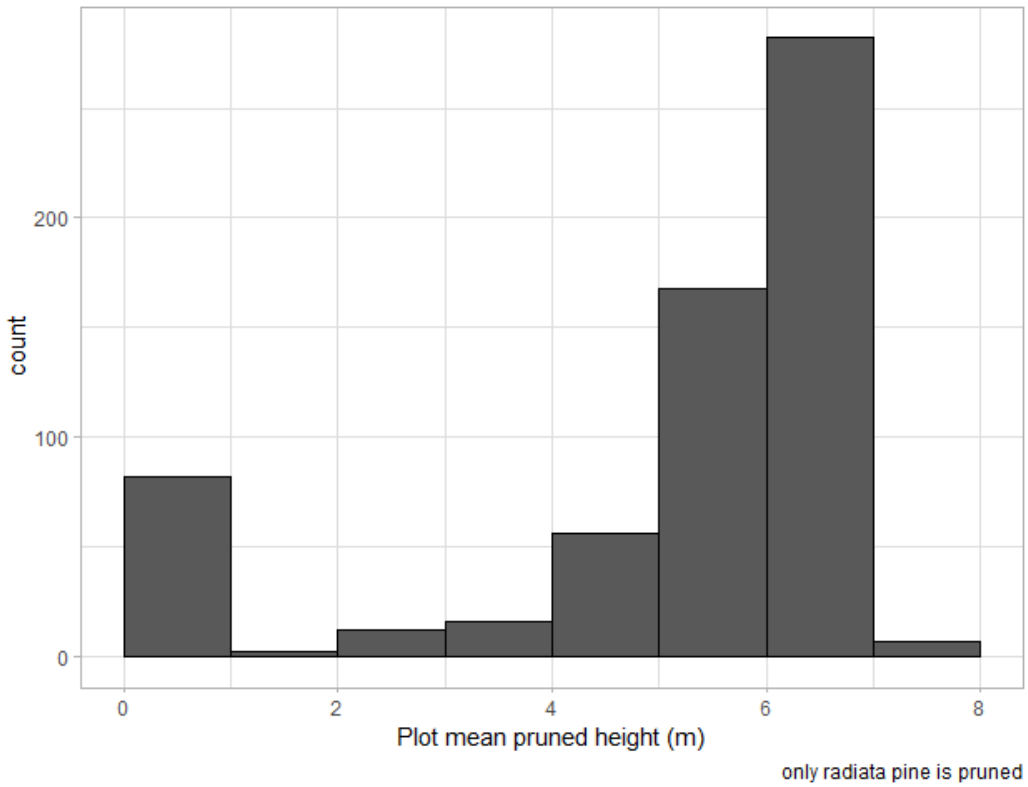

Figure 6-44: Distribution of plot pruned height. PRAD = radiata pine.

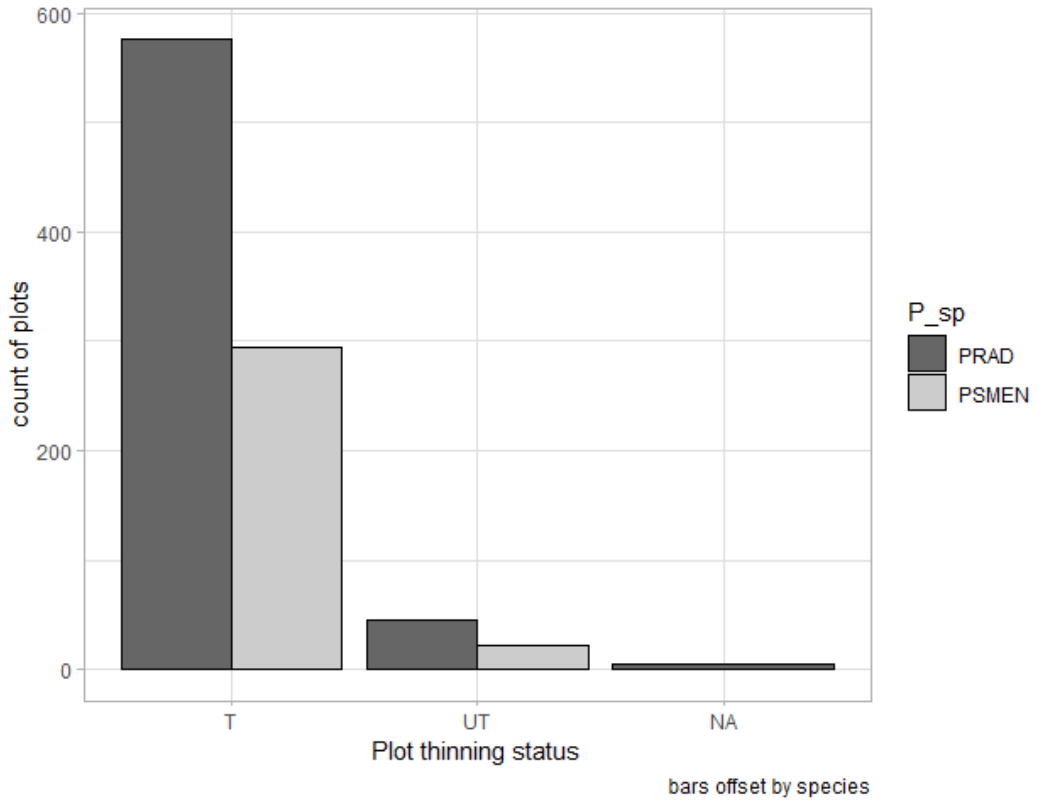

Figure 6-45: Plot thinned/unthinned status by species.

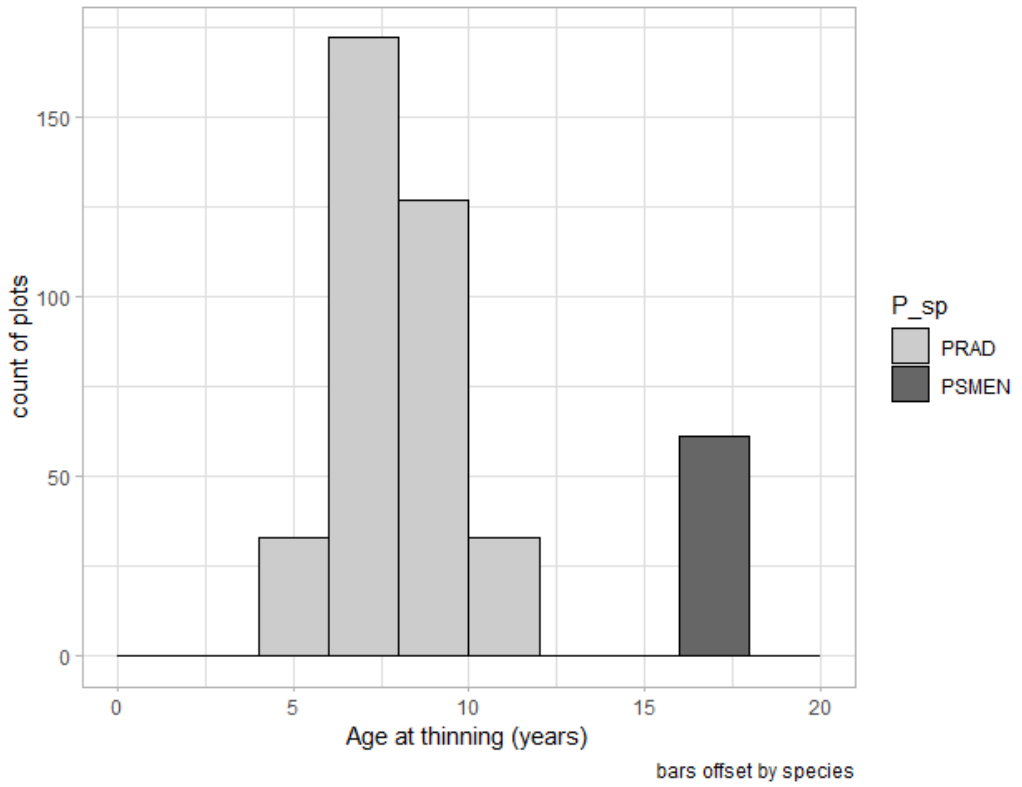

Figure 6-46: Distribution of plot age at thinning (Age_thin).

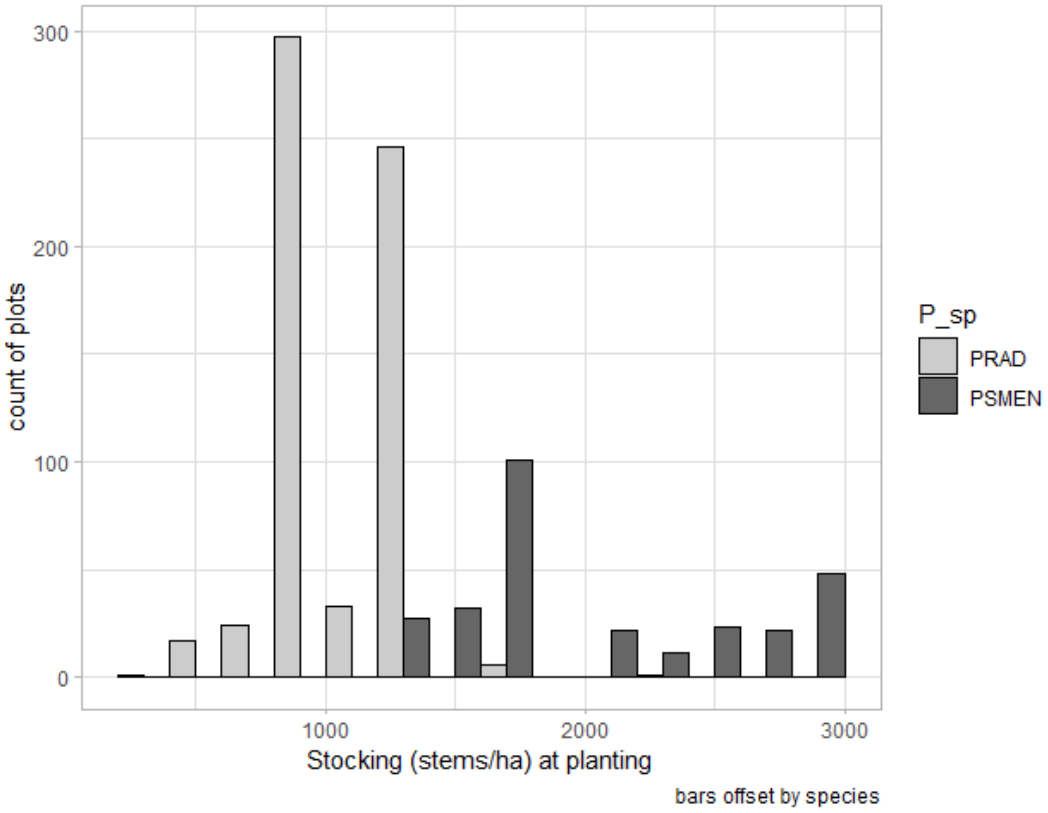

Figure 6-47: Distribution of stocking at time of planting (Estab_sph).

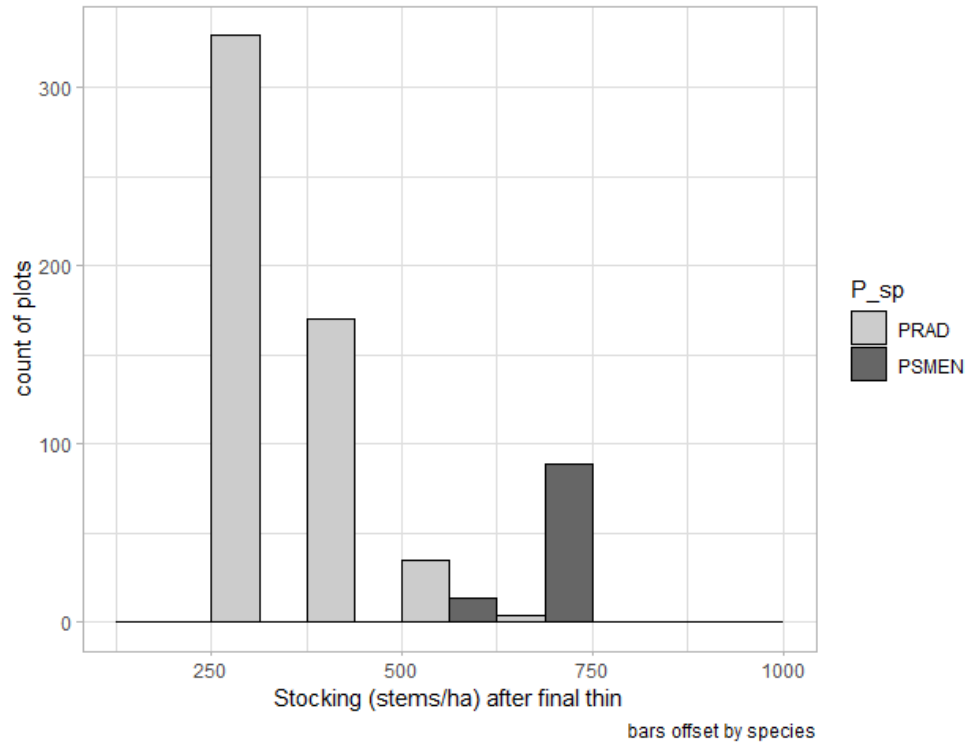

Figure 6-48: Distribution of stocking after final thinning (Final_sph).

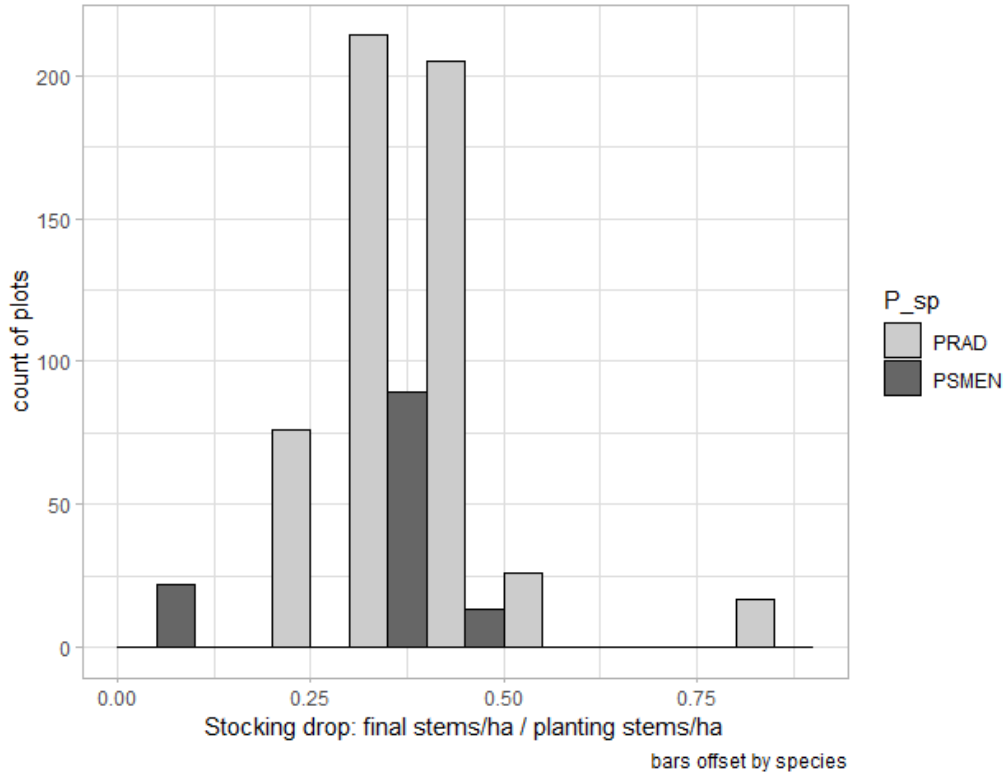

Figure 6-49: Distribution of proportion drop in stocking between planting and final thinning (Sph_drop).

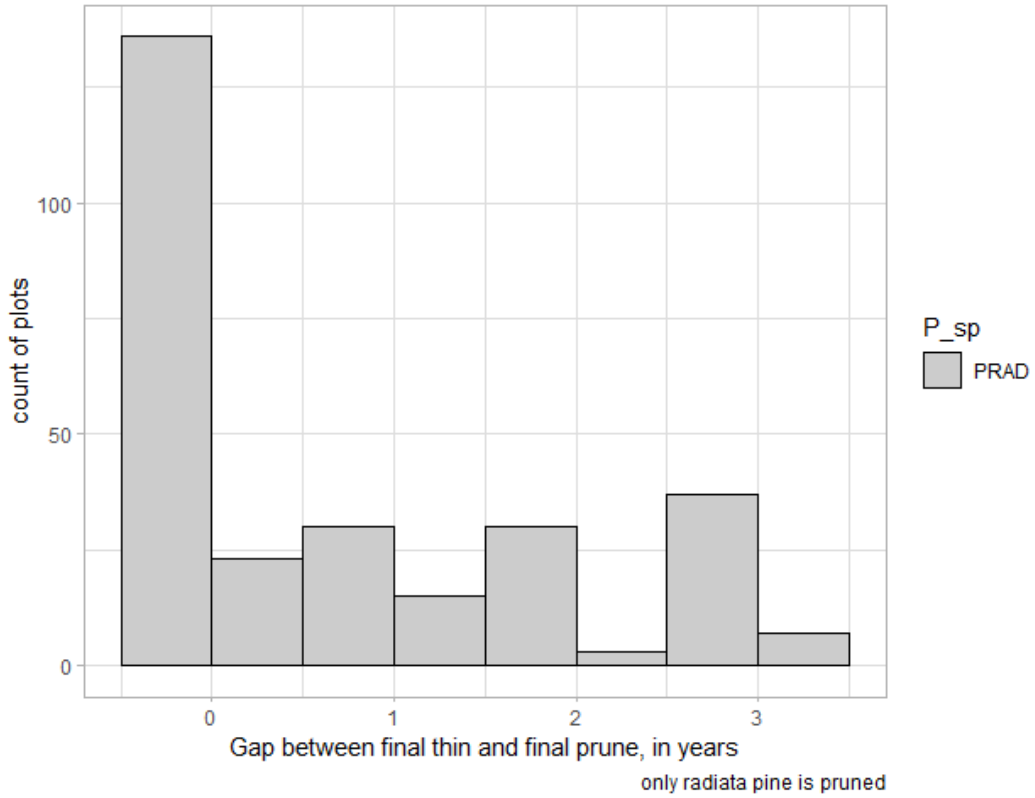

Figure 6-50: Distribution of gap between final pruning and final thinning in years (T_P_gap).

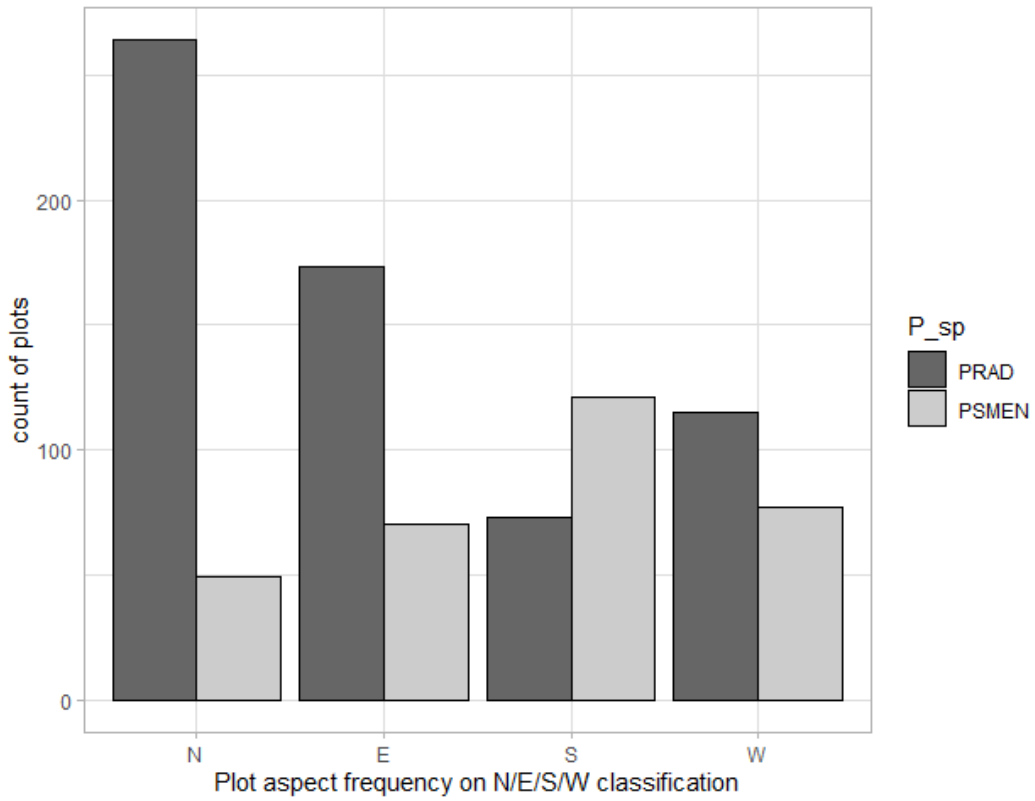

Figure 6-51: Frequency of plot predominant aspect by north/south/east/west classification (card_4wayN).

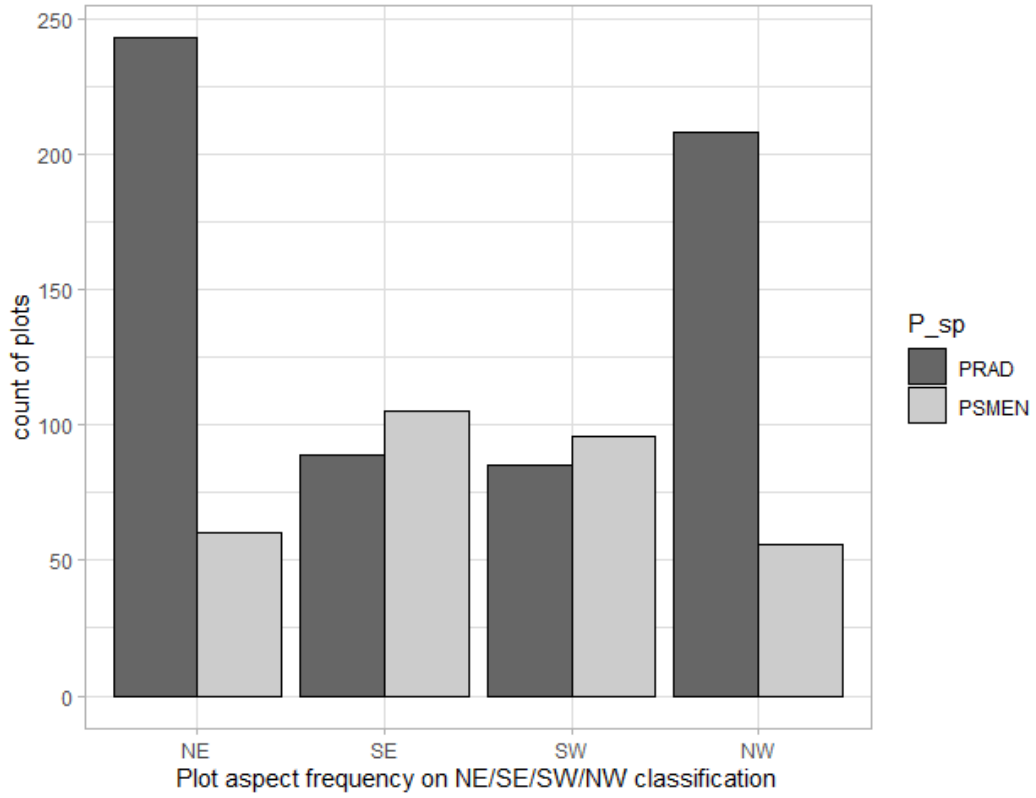

Figure 6-52: Frequency of plot predominant aspect by north-east/south-east/south-west/north-west classification (card_4wayNE).

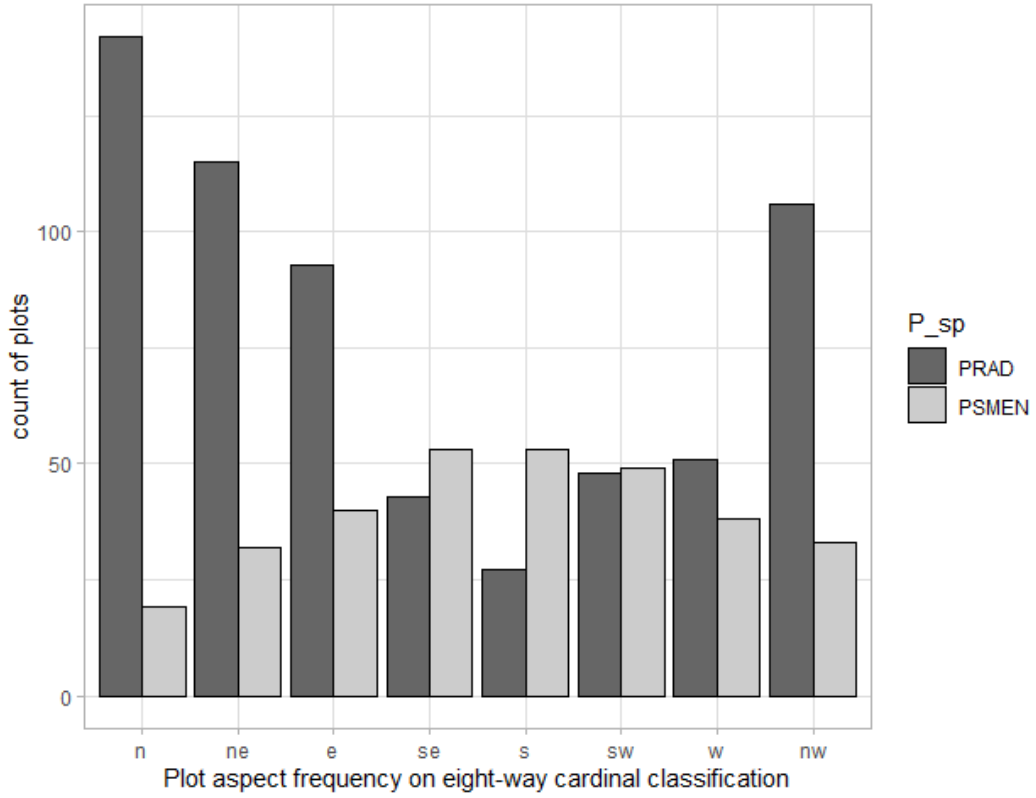

Figure 6-53: Frequency of plot predominant aspect by north/north-east/east/south-east/south/south-west/west/north-west classification (card_8way).

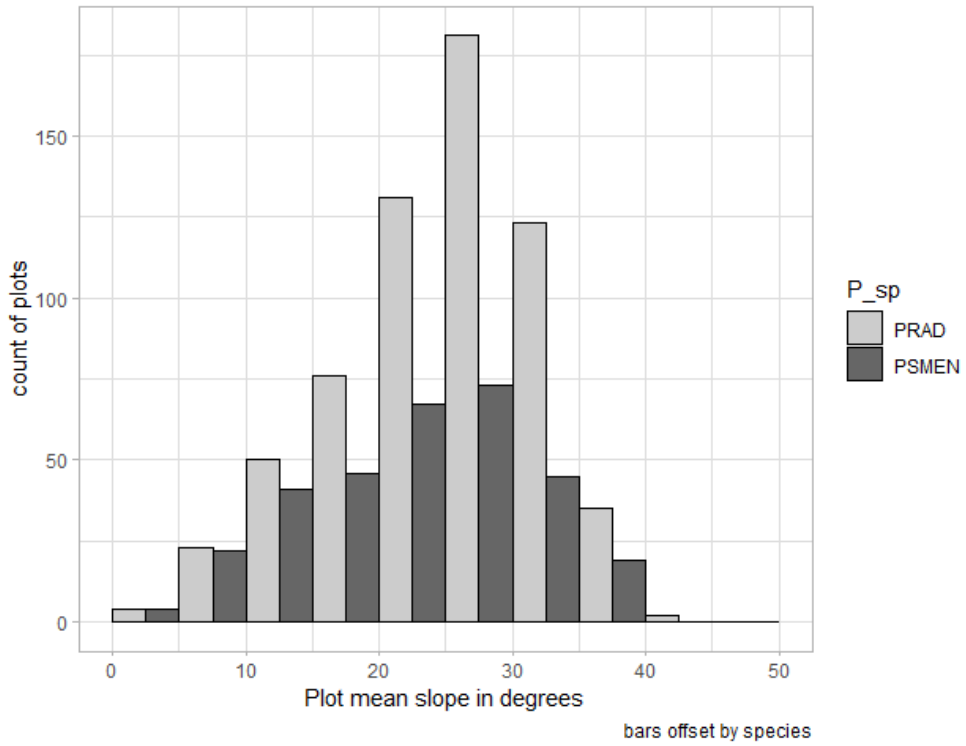

Figure 6-54: Distribution of plot mean slope (P_slope).

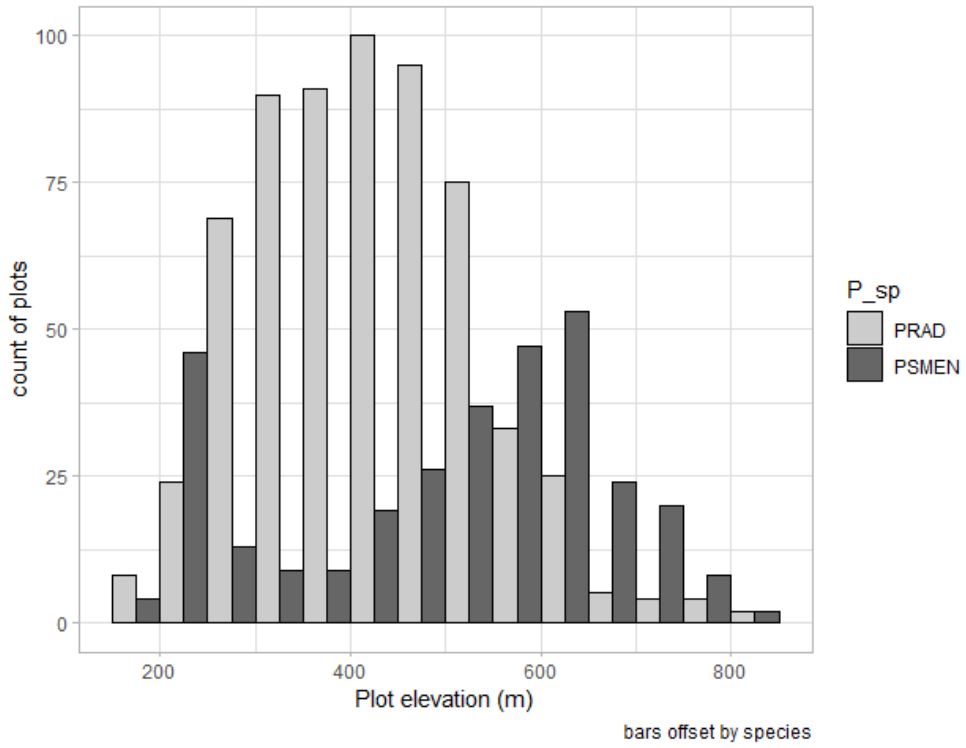

Figure 6-55: Distribution of plot mean elevation (P_alt).

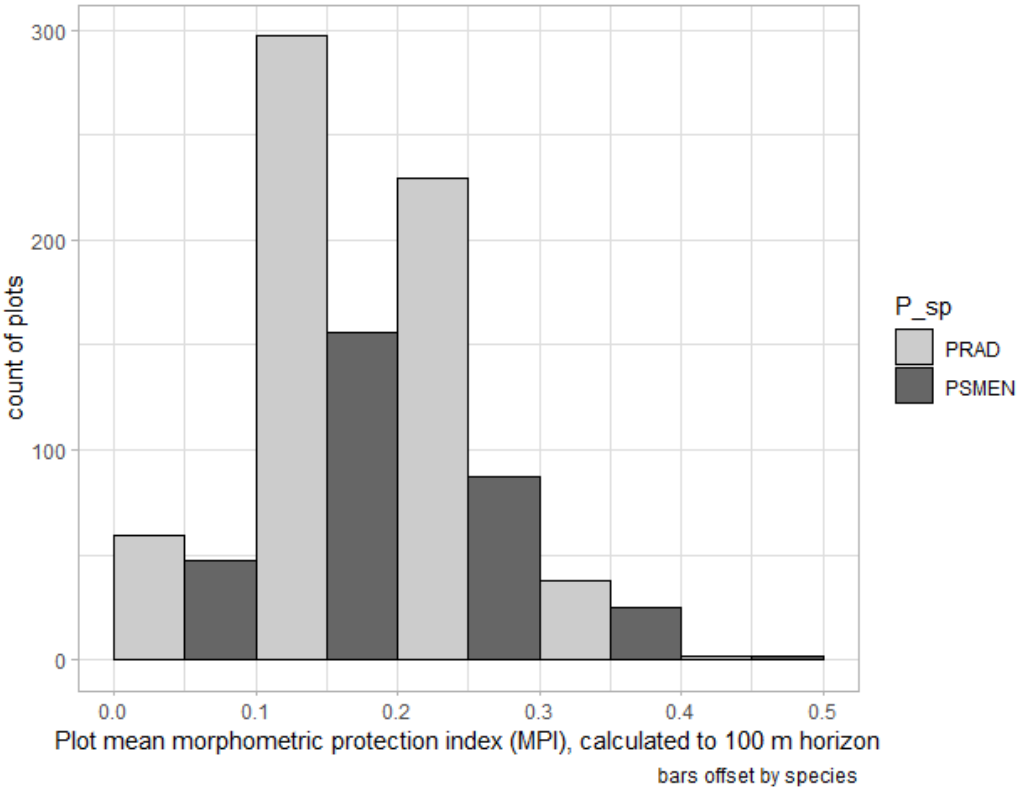

Figure 6-56: Distribution of plot mean morphometric protection index at a 100 m horizon (MPI_100). 0 = completely sheltered, 1 = completely exposed.

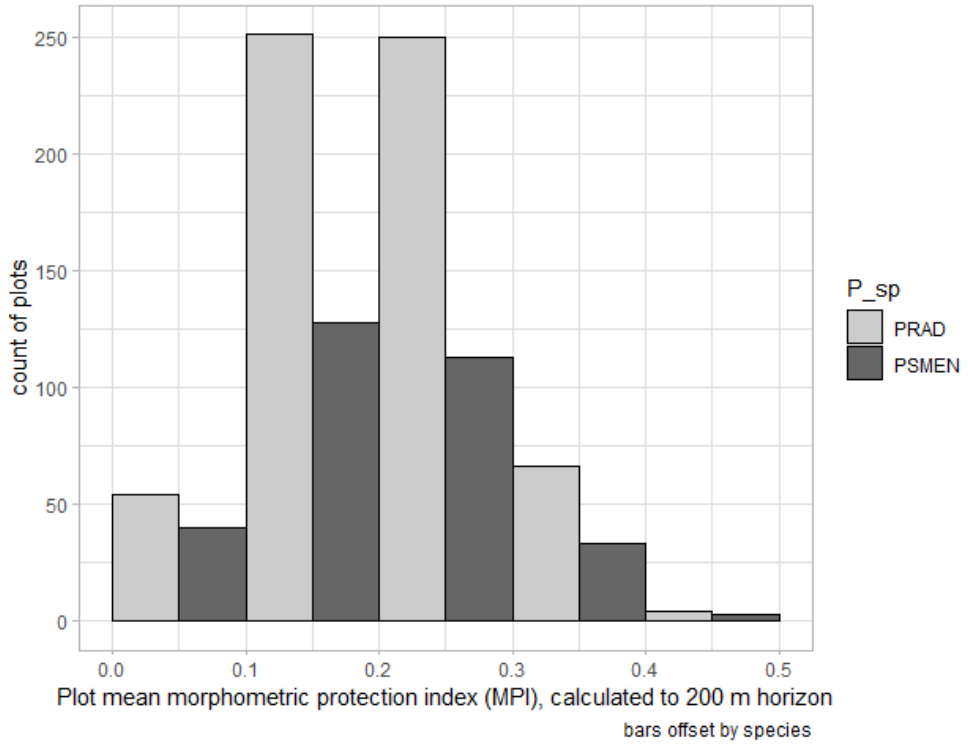

Figure 6-57: Distribution of plot mean morphometric protection index at a 200 m horizon (MPI_200). 0 = completely sheltered, 1 = completely exposed.

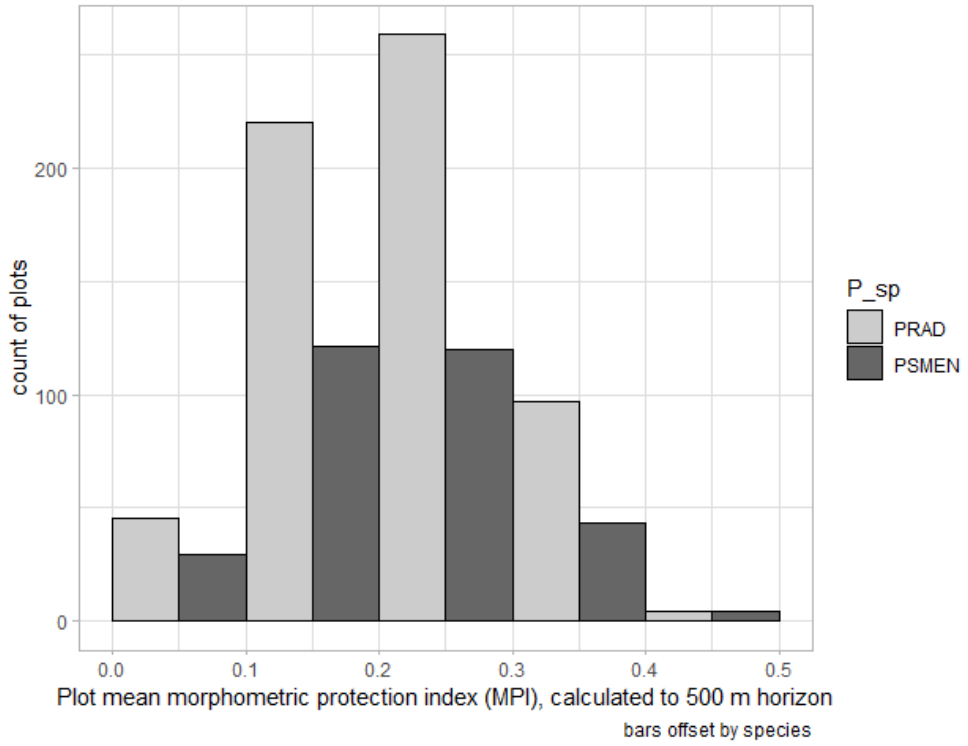

Figure 6-58: Distribution of plot mean morphometric protection index at a 500 m horizon (MPI_500). 0 = completely sheltered, 1 = completely exposed.

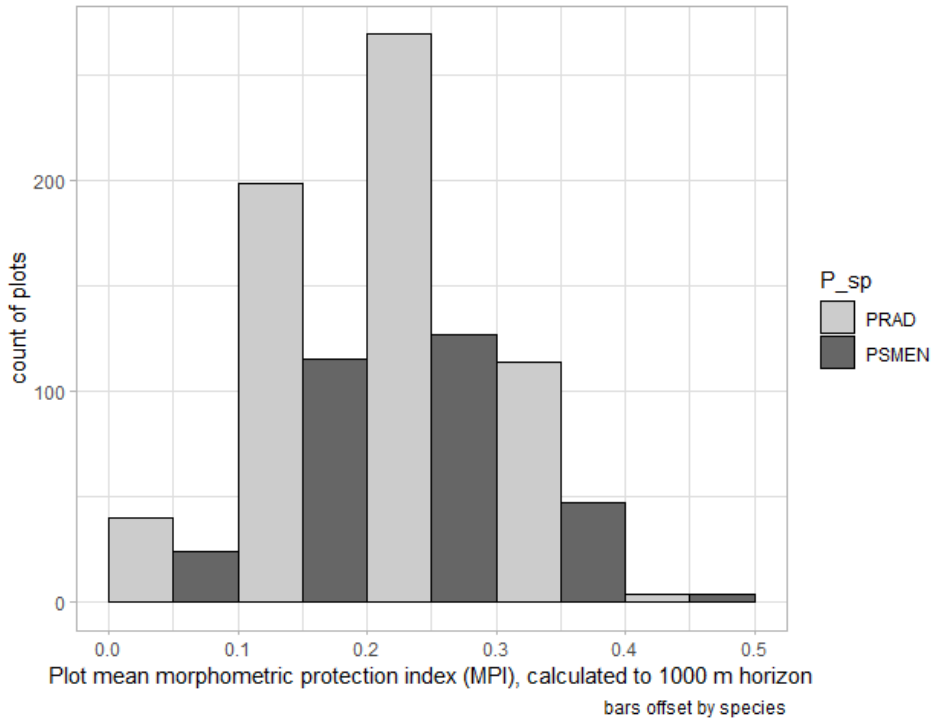

Figure 6-59: Distribution of plot mean morphometric protection index at a 1000 m horizon (MPI_1000). 0 = completely sheltered, 1 = completely exposed.

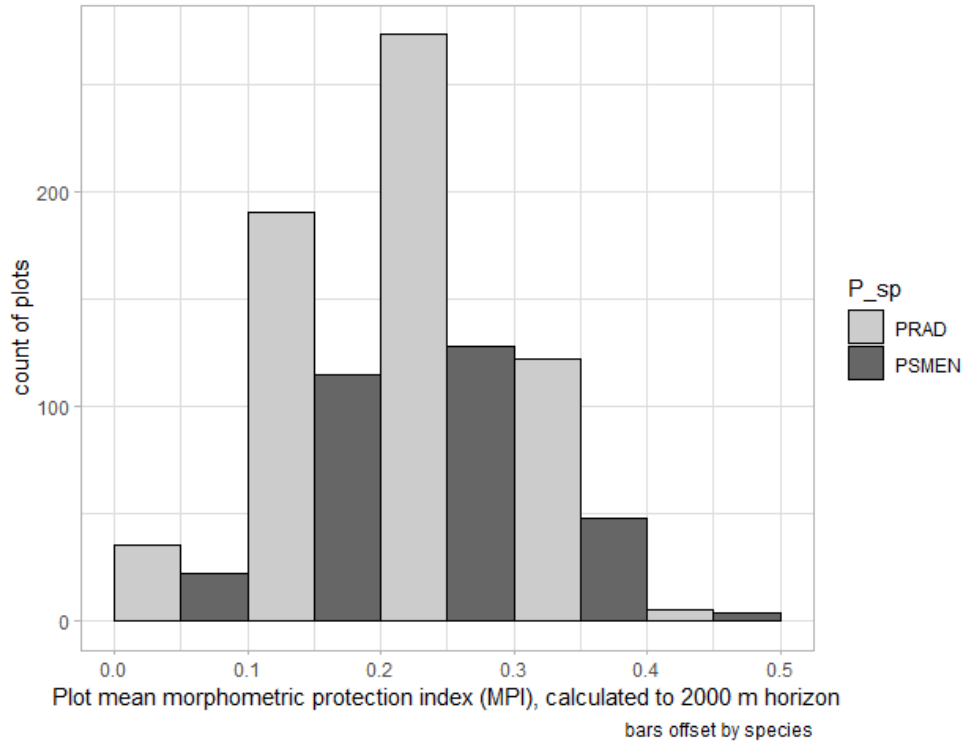

Figure 6-60: Distribution of plot mean morphometric protection index at a 2000 m horizon (MPI_2000). 0 = completely sheltered, 1 = completely exposed.

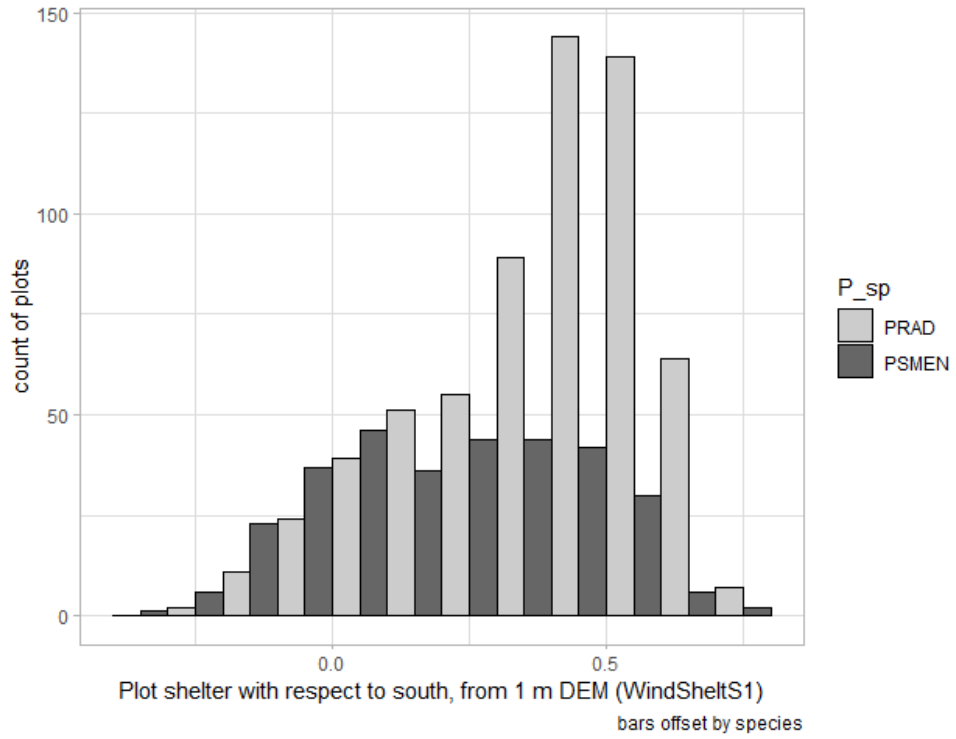

Figure 6-61: Distribution of plot shelter with respect to south (WindSheltS1). Values less than zero are wind-shadowed, values greater than zero are wind-exposed.

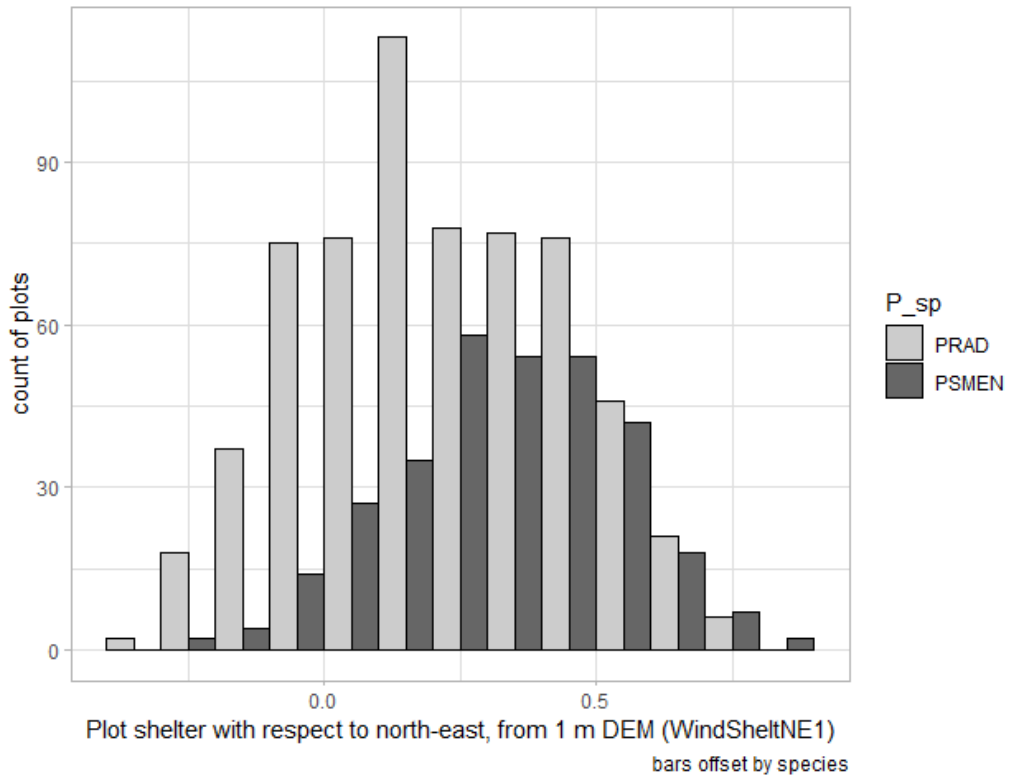

Figure 6-62: Distribution of plot shelter with respect to north-east (WindSheltNE1). Values less than zero are wind-shadowed, values greater than zero are wind-exposed.

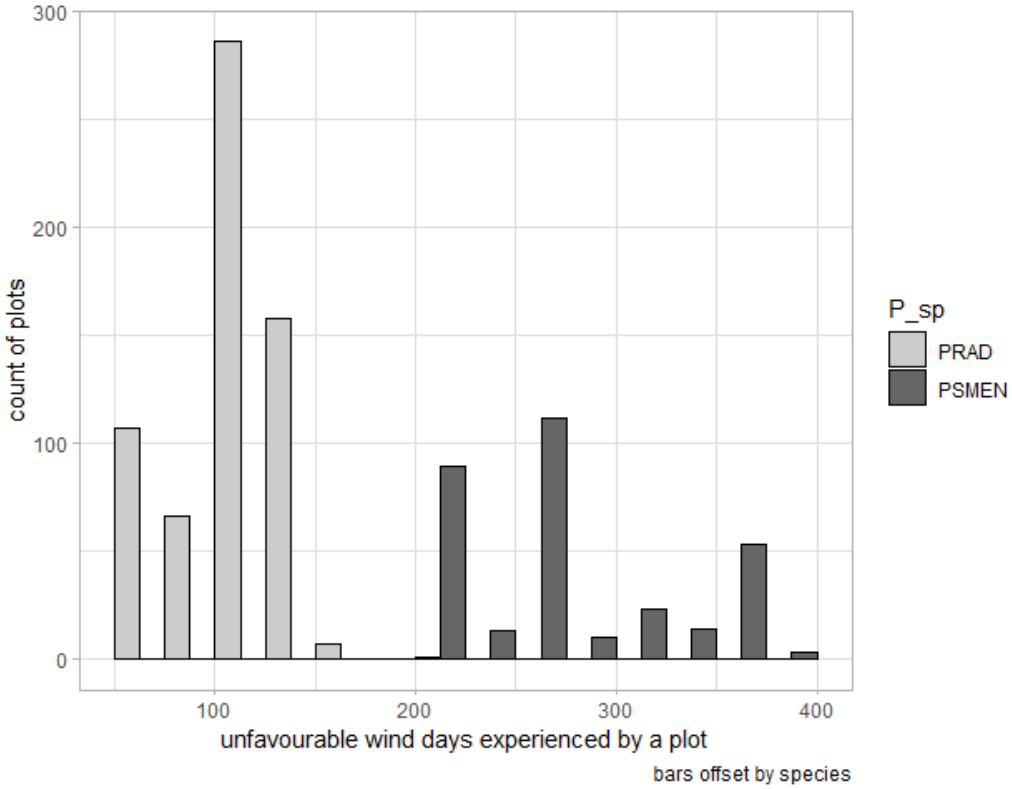

Figure 6-63: Distribution of unfavourable wind days experienced per plot (u_wind_tim).

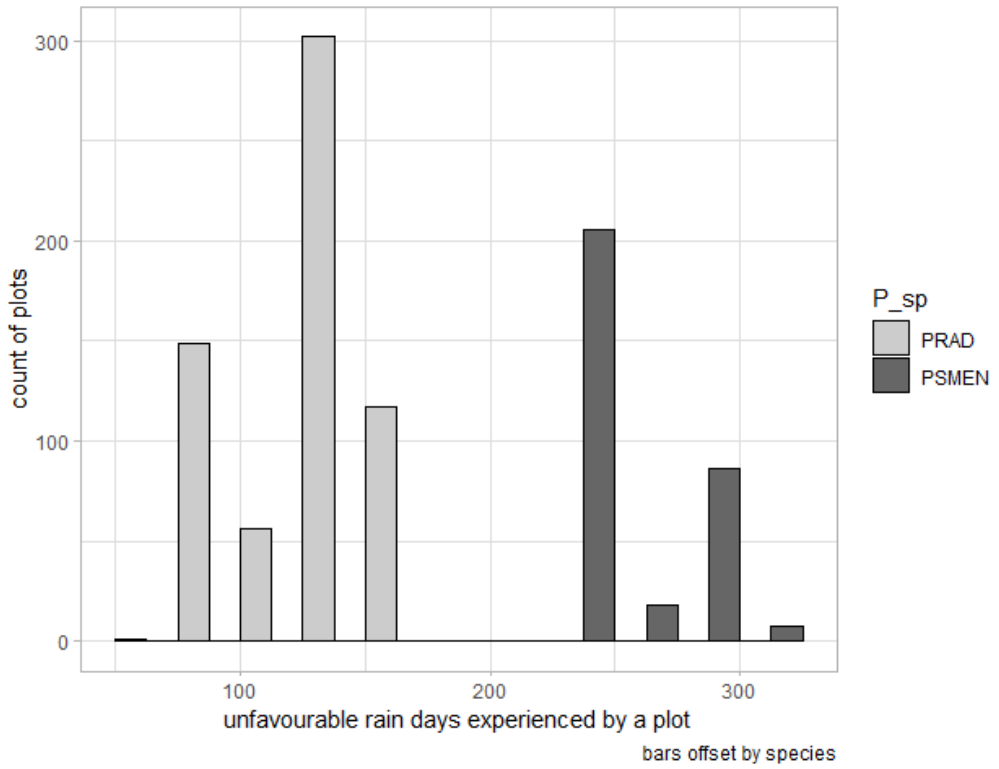

Figure 6-64: Distribution of unfavourable rain days experienced per plot (u_rain).

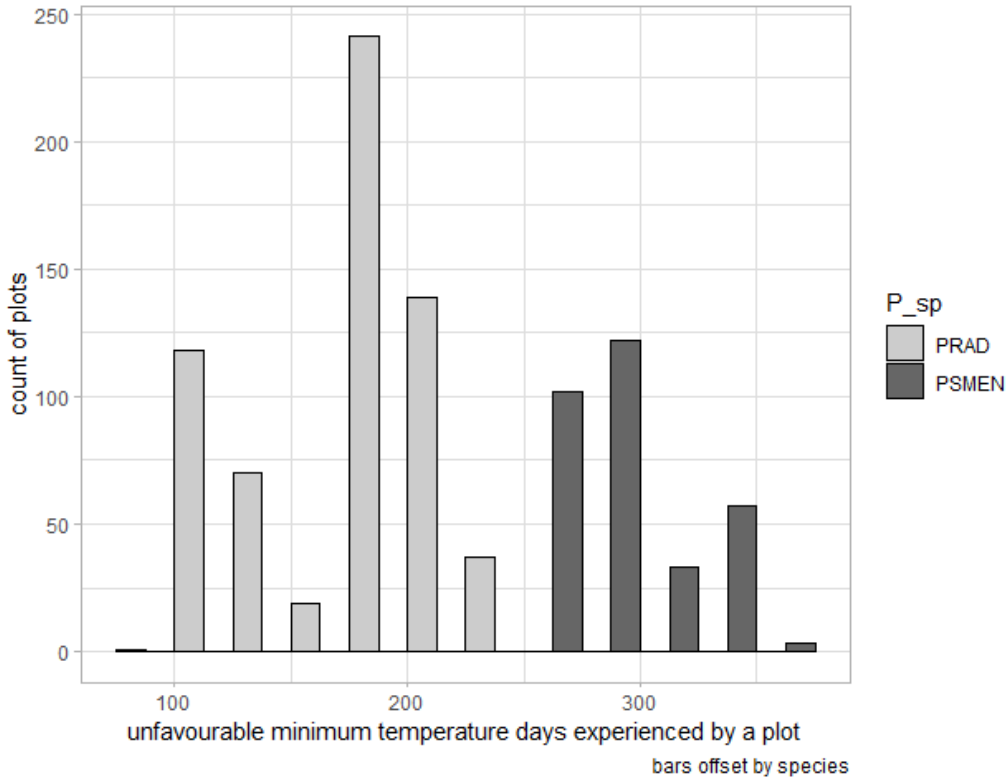

Figure 6-65: Distribution of unfavourable minimum temperature days experienced per plot (u_min_temp).

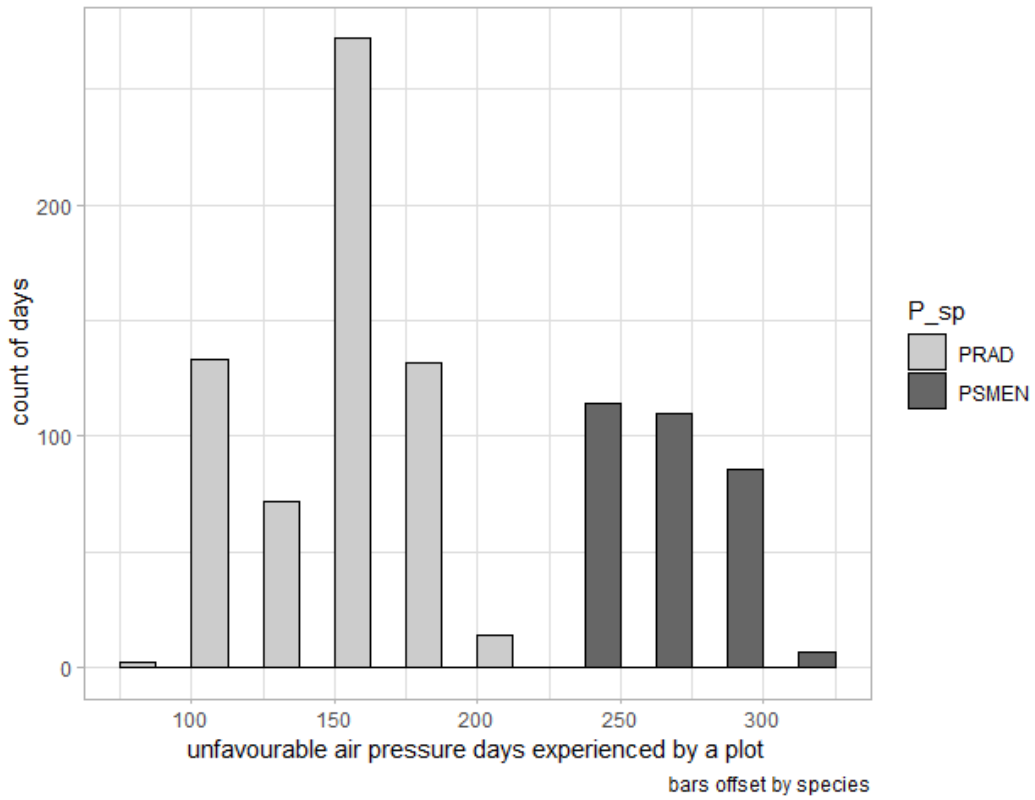

Figure 6-66: Distribution of unfavourable air (barometric) pressure days experienced per plot (u_air_pr).

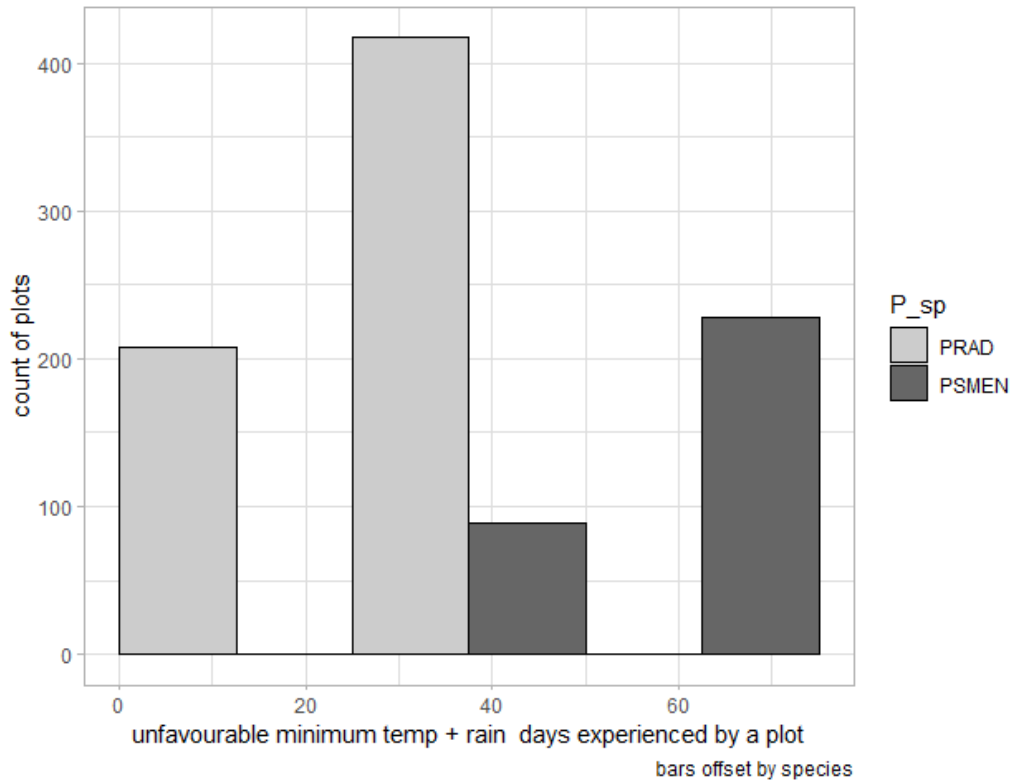

Figure 6-67: Distribution of unfavourable minimum temperature and rain days experienced per plot (u_mint_rain).

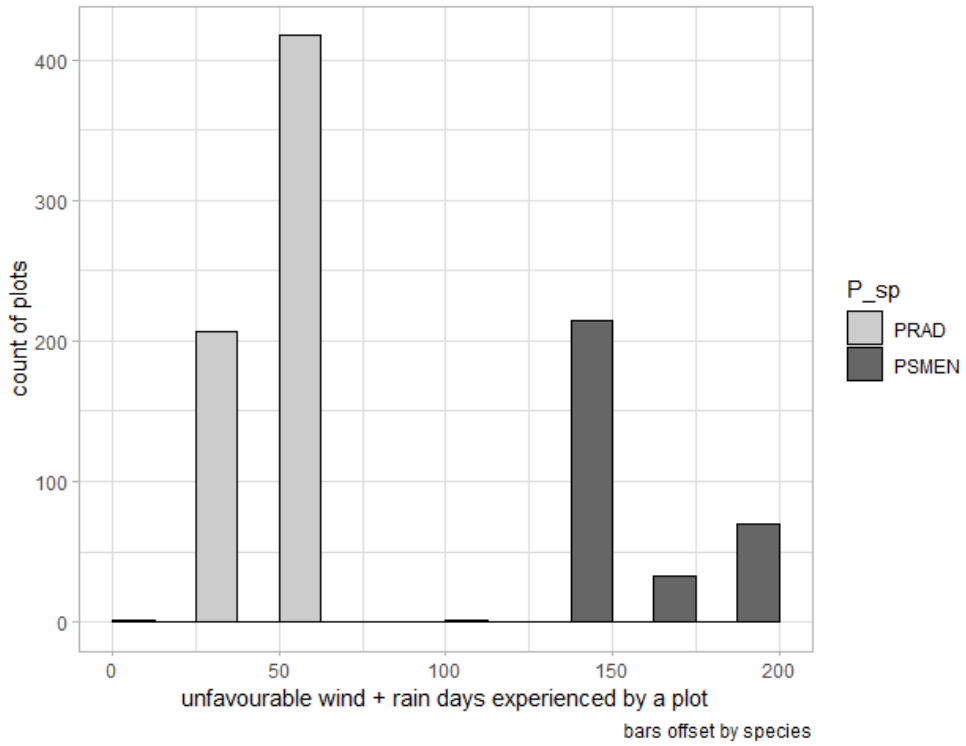

Figure 6-68: Distribution of unfavourable wind and rain days experienced per plot (u_rain_wind_tim).

6.5.3 Correlations between explanatory variables

In this section, the size and colour depth of the dots in the figures represents the absolute value of the correlation between response variables and explanatory variables. The colour represents whether the correlation is positive or negative. A colour legend is included in each figure.

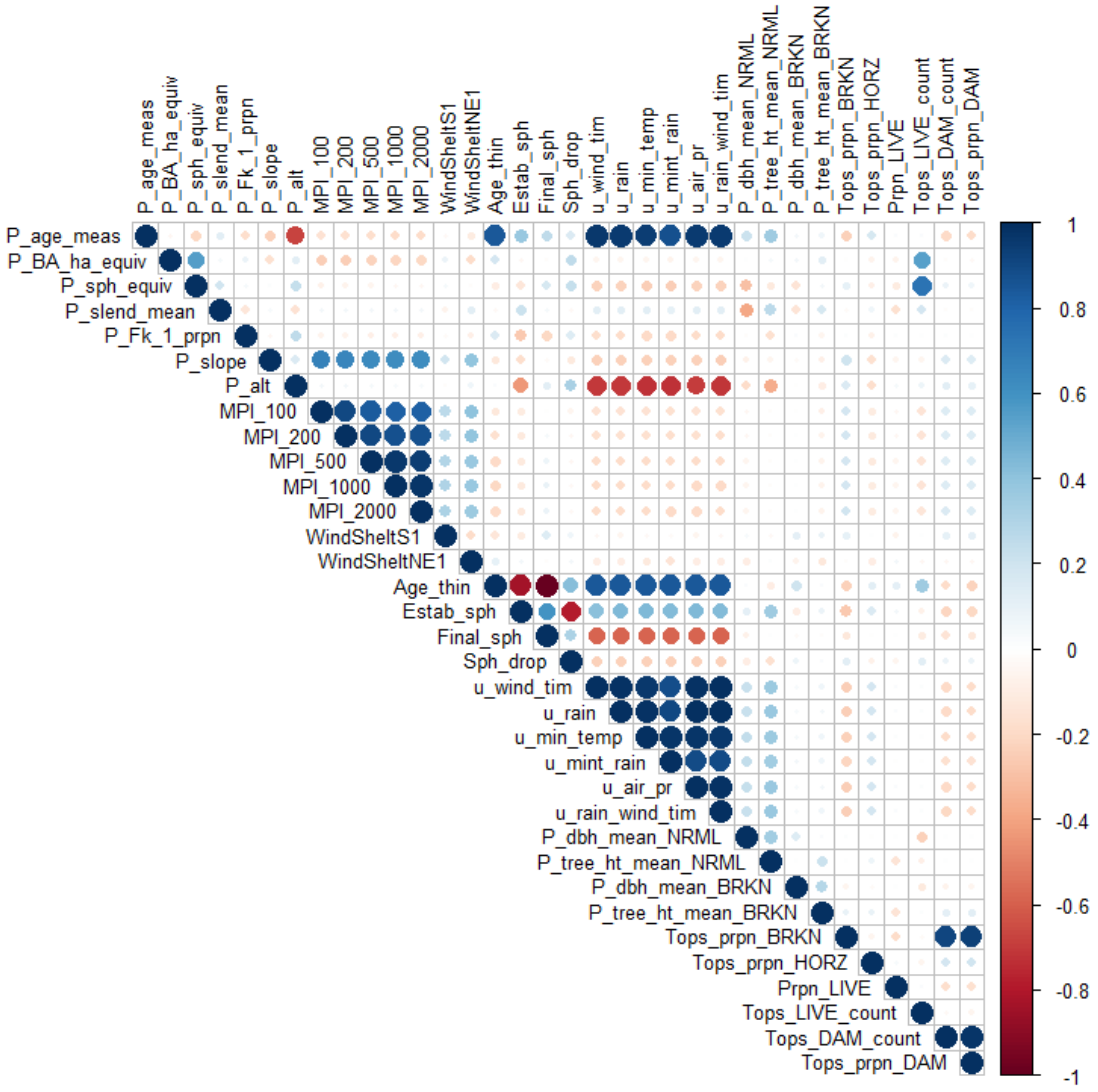

Figure 6-69: correlations between explanatory variables for radiata pine.

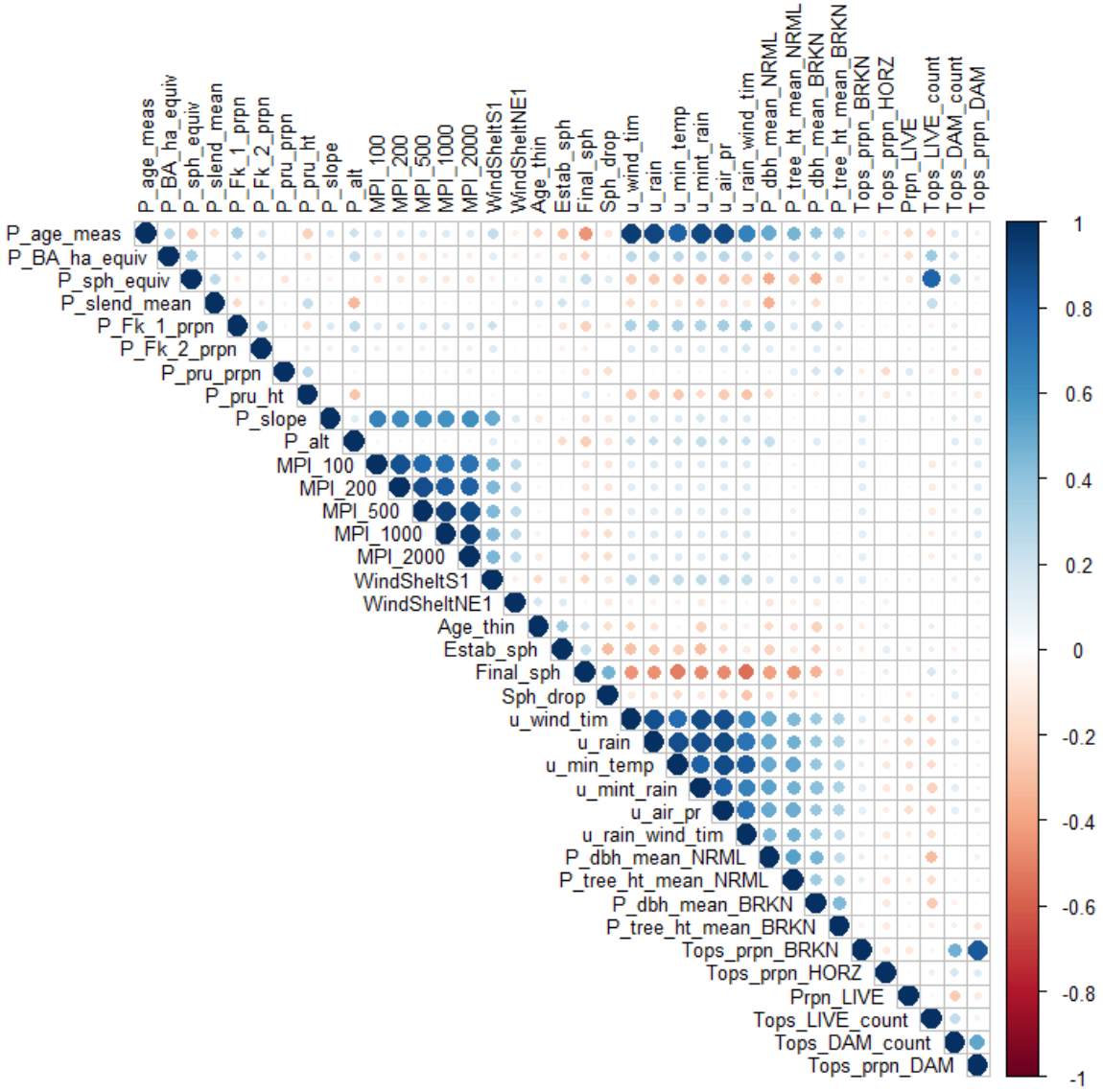

Figure 6-70: correlations between explanatory variables for radiata pine.

6.5.4 Correlations between response variables and explanatory variables

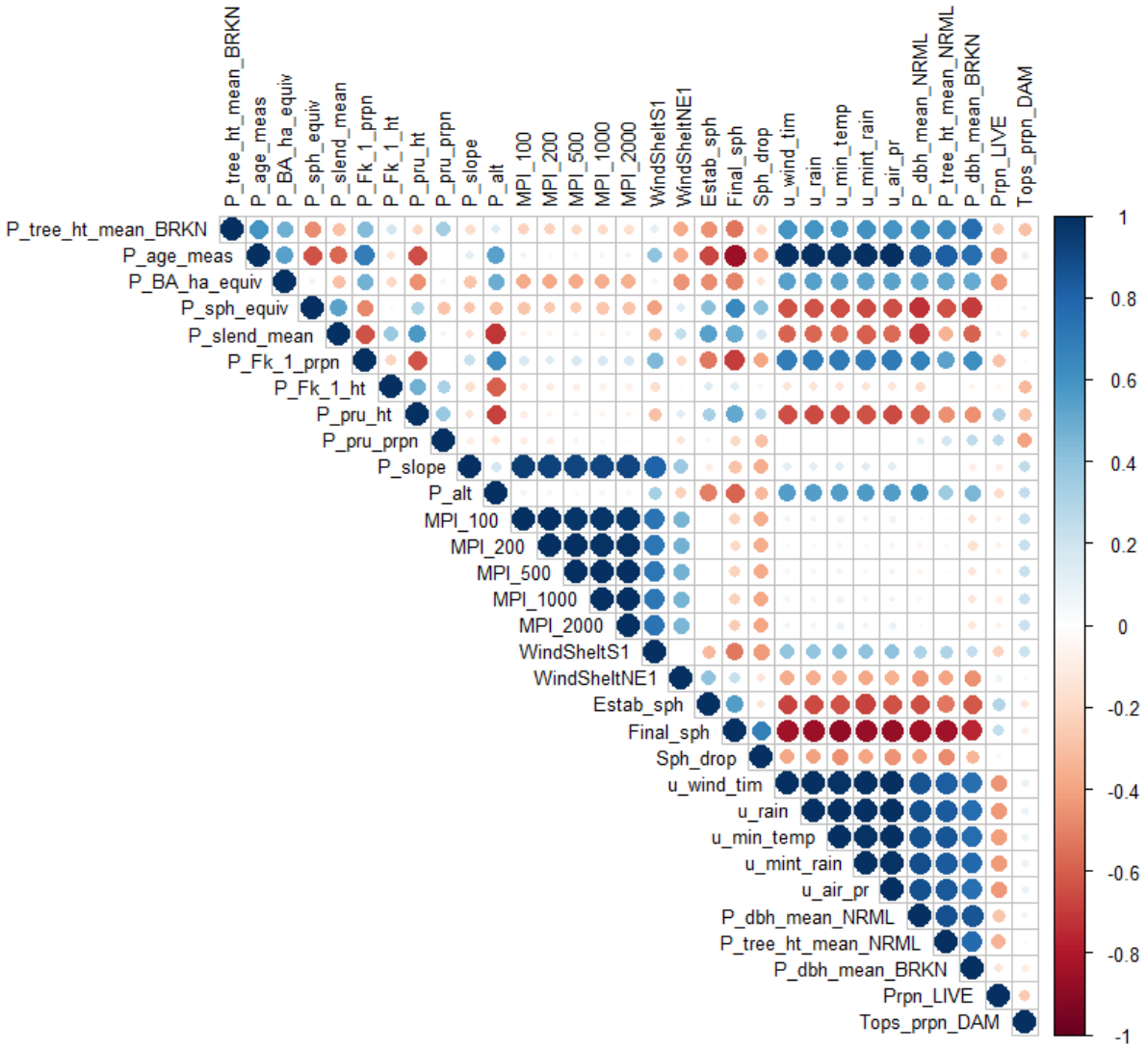

Figure 6-71: correlations between response variable P_tree_ht_mean_BRKN and explanatory variables for Douglas-fir.

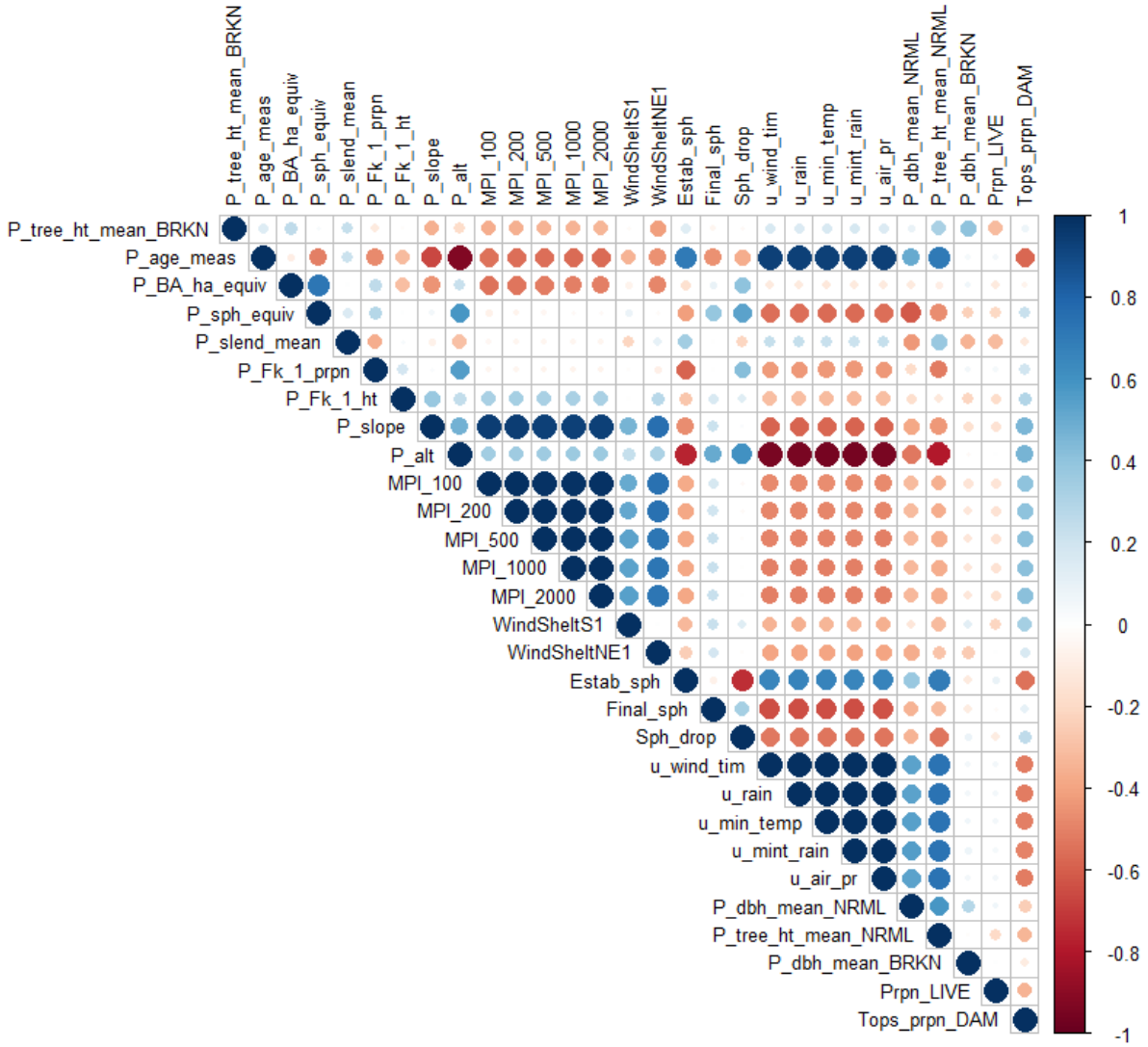

Figure 6-72: correlations between response variable P_tree_ht_mean_BRKN and explanatory variables for Douglas-fir.

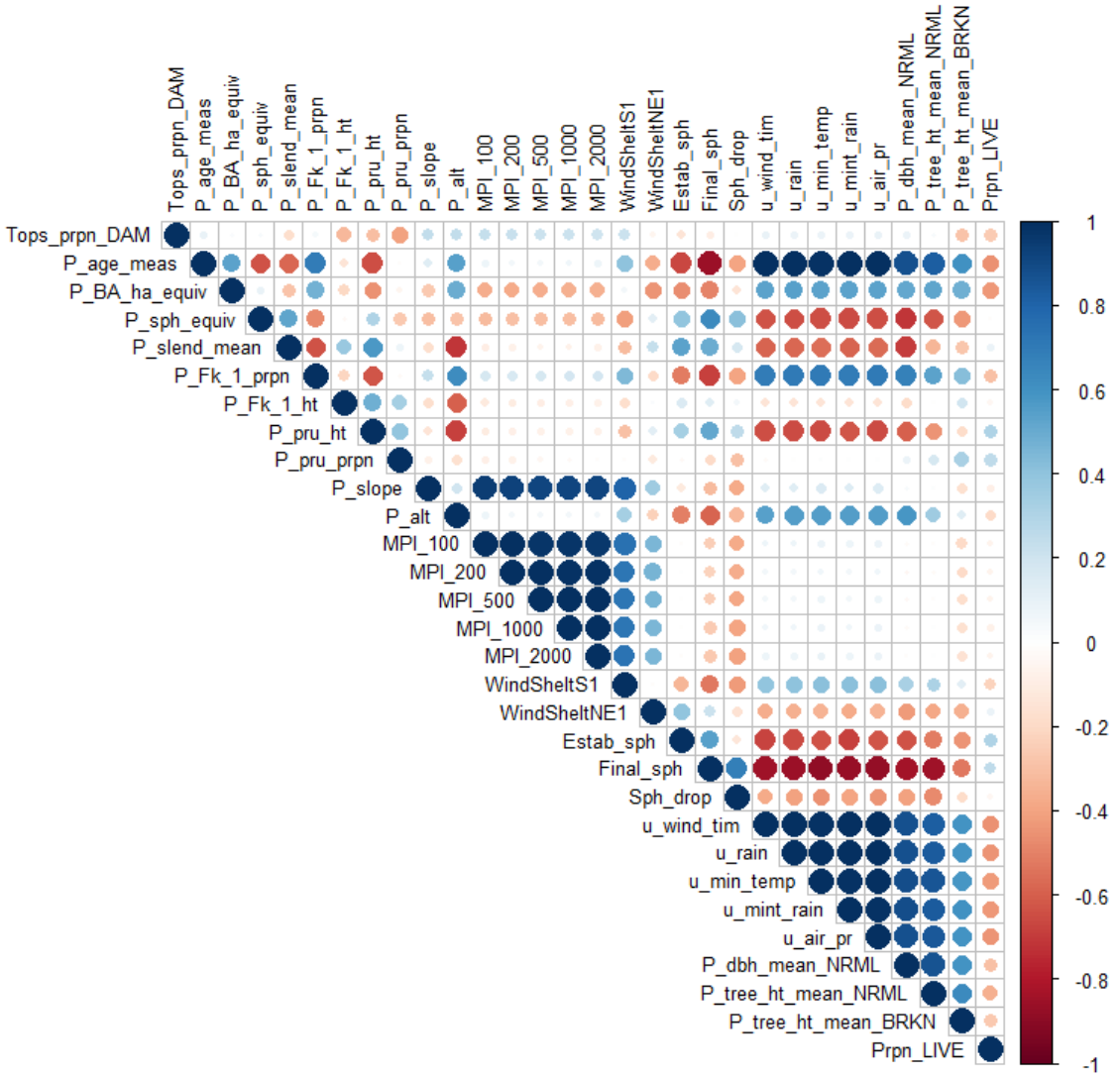

Figure 6-73: correlations between response variable Tops_prpn_DAM and explanatory variables, for all plots, for radiata pine.

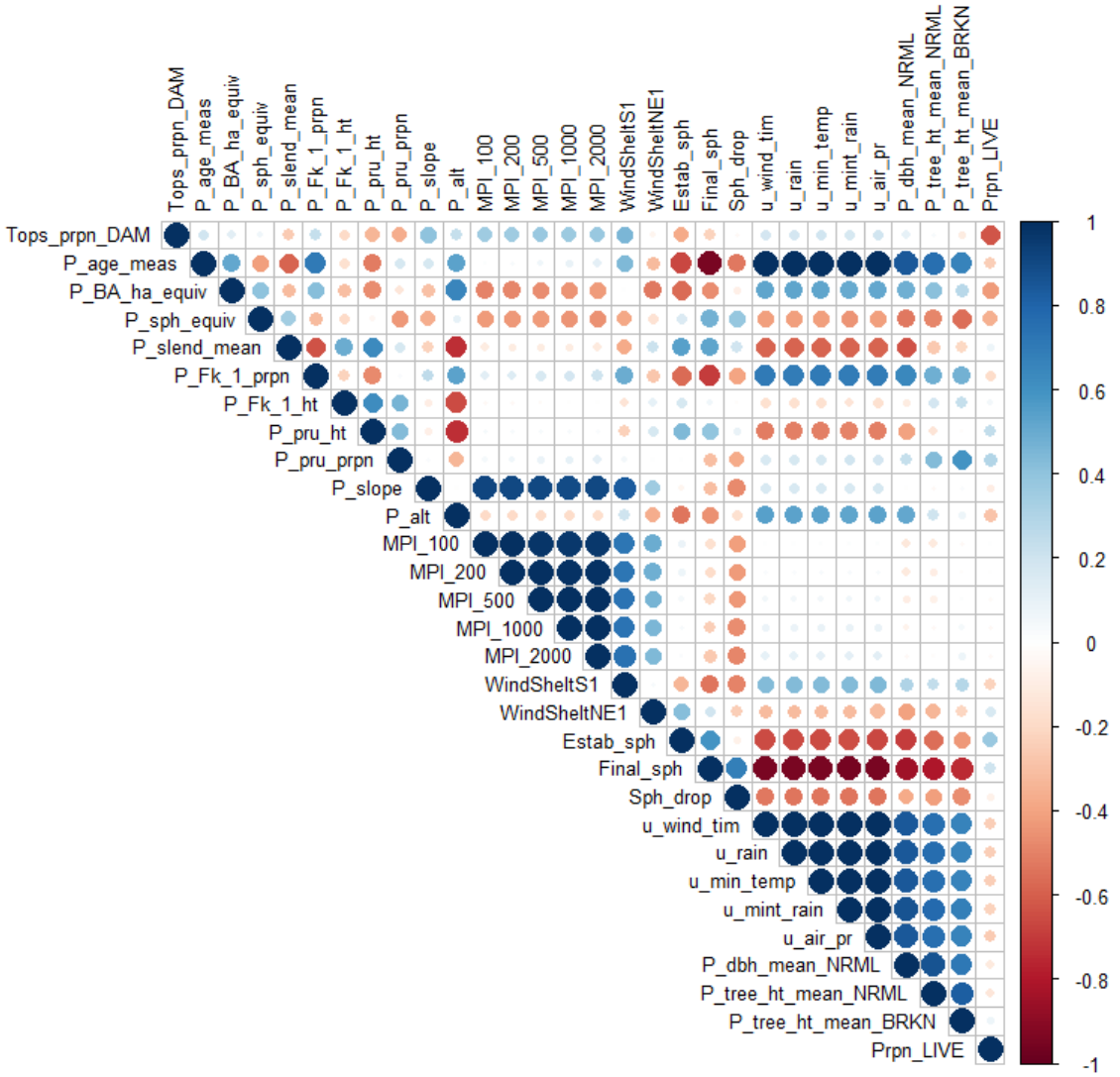

Figure 6-74: correlations between Tops_prpn_DAM and explanatory variables, for plots with all tops assessed, for radiata pine.

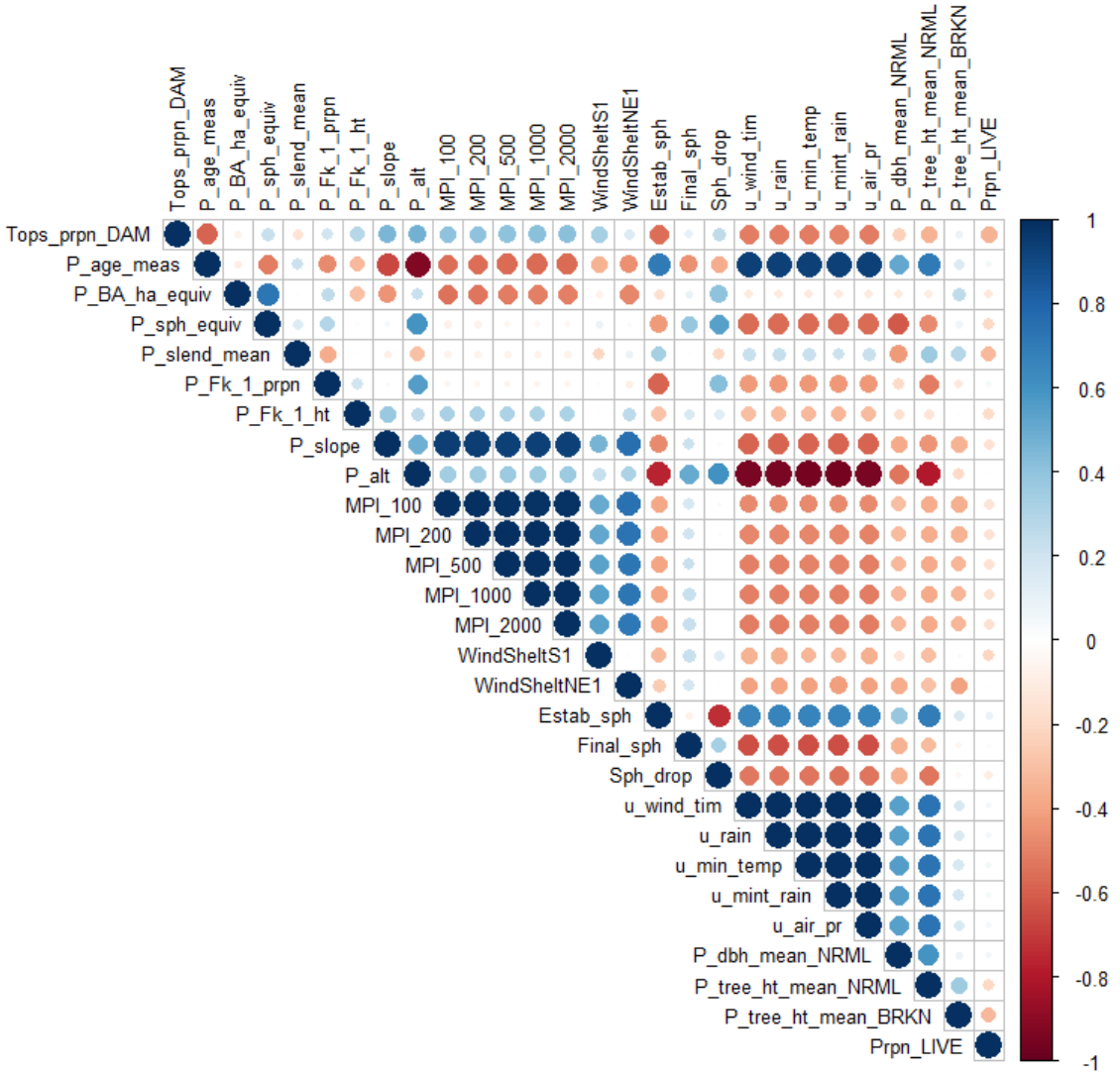

Figure 6-75: correlations between Tops_prpn_DAM and explanatory variables, for plots with all tops assessed, for Douglas-fir.

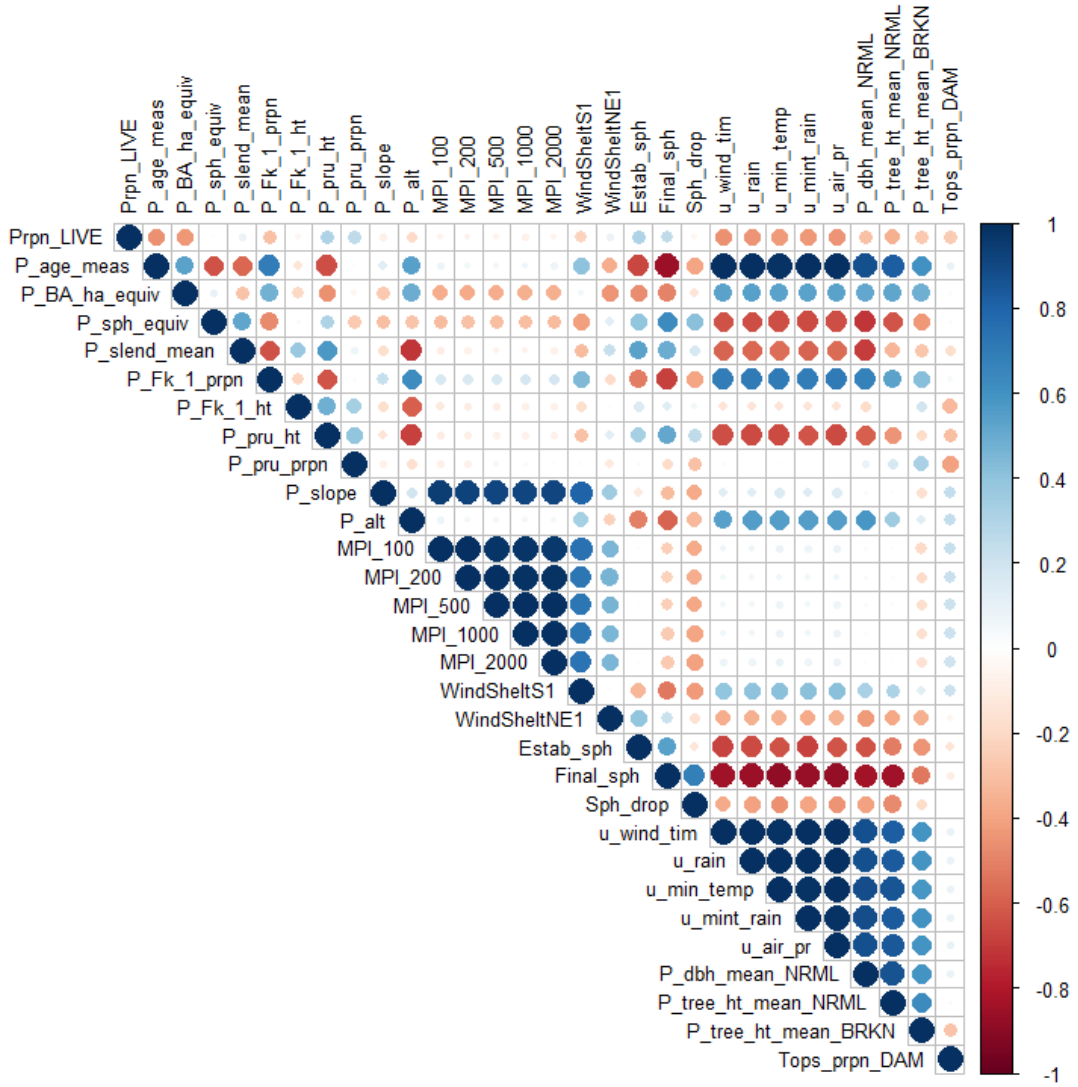

Figure 6-76: correlations between response variable Prpn_LIVE and explanatory variables, radiata pine.

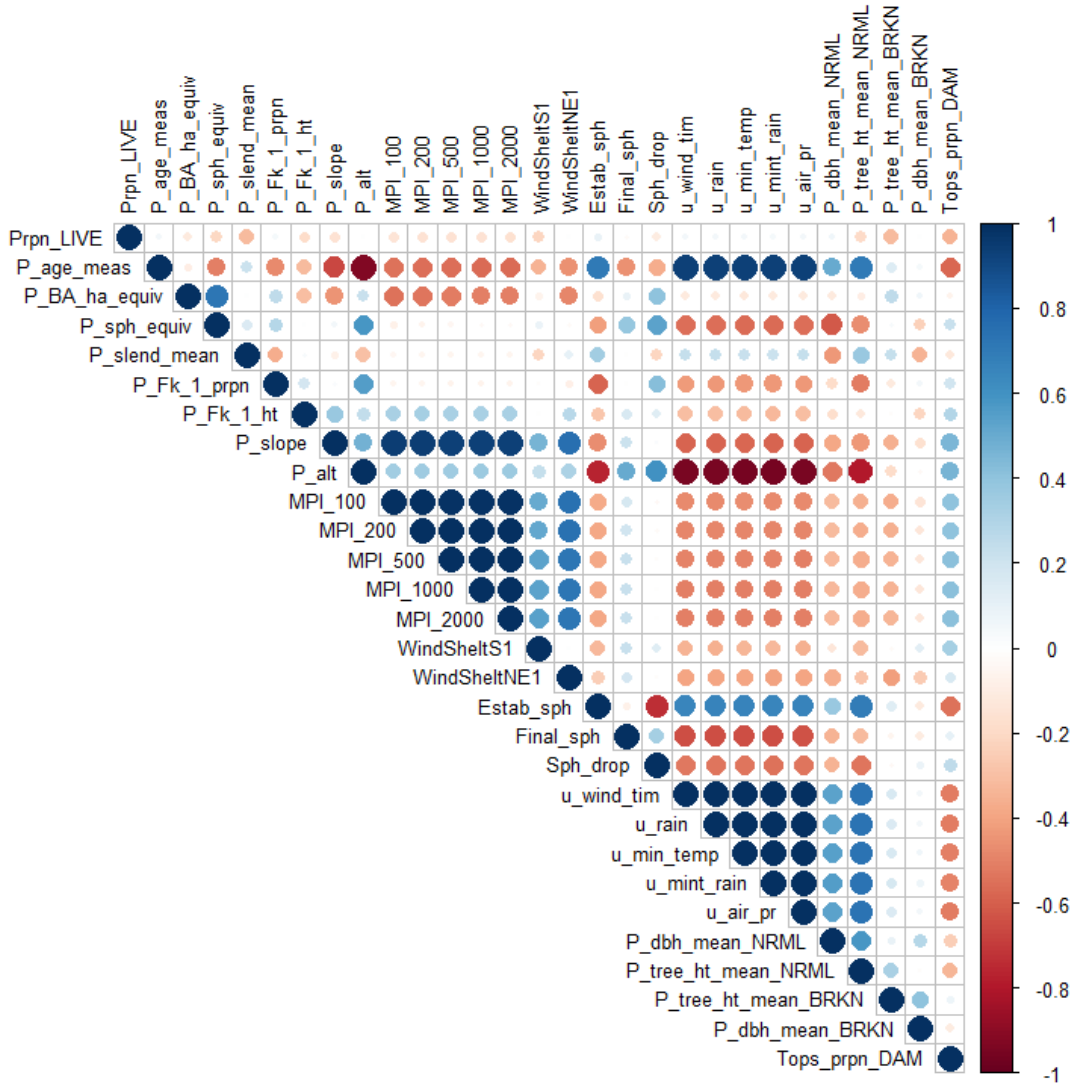

Figure 6-77: correlations between response variable Prpn_LIVE and explanatory variables, for Douglas-fir.

6.5.5 Classification and regression trees

The following figures are classification and regression trees (CARTs), for each combination of response variable (*P_tree_ht_mean_BRKN*, *Tops_prpn_DAM*, and *Prpn_LIVE*) and species (radiata pine and Douglas-fir). As these are exploratory data analysis, all data are included, i.e. the dataset is not split to validation and test sets. Variable *P_stand* could not be included, because it has 85 levels, and factor variables can have only up to 32 variables when used with function *tree*. Variables with high occurrences of NA were also excluded, for the reasons described in section 2.6.

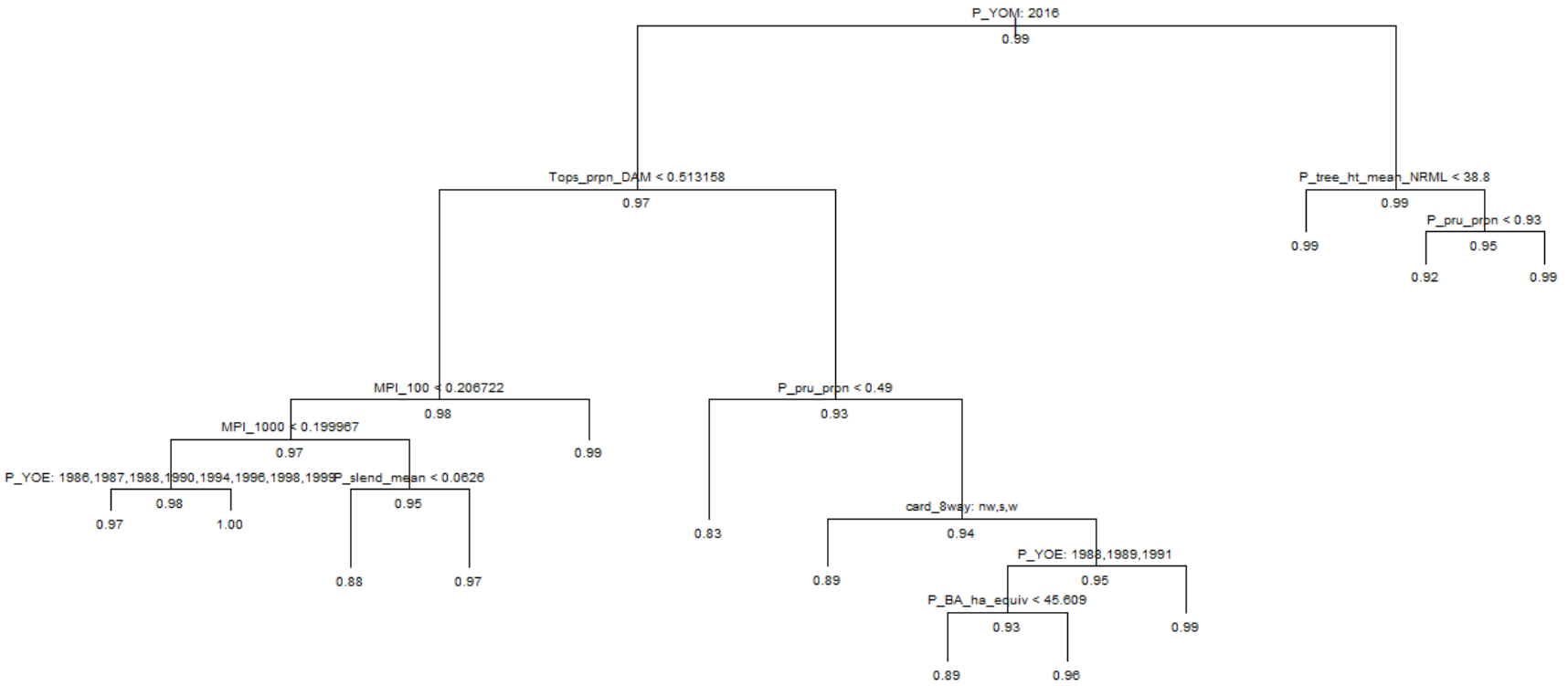

Figure 6-78: CART for response variable P_tree_ht_mean_BRKN and explanatory variables, for radiata pine.

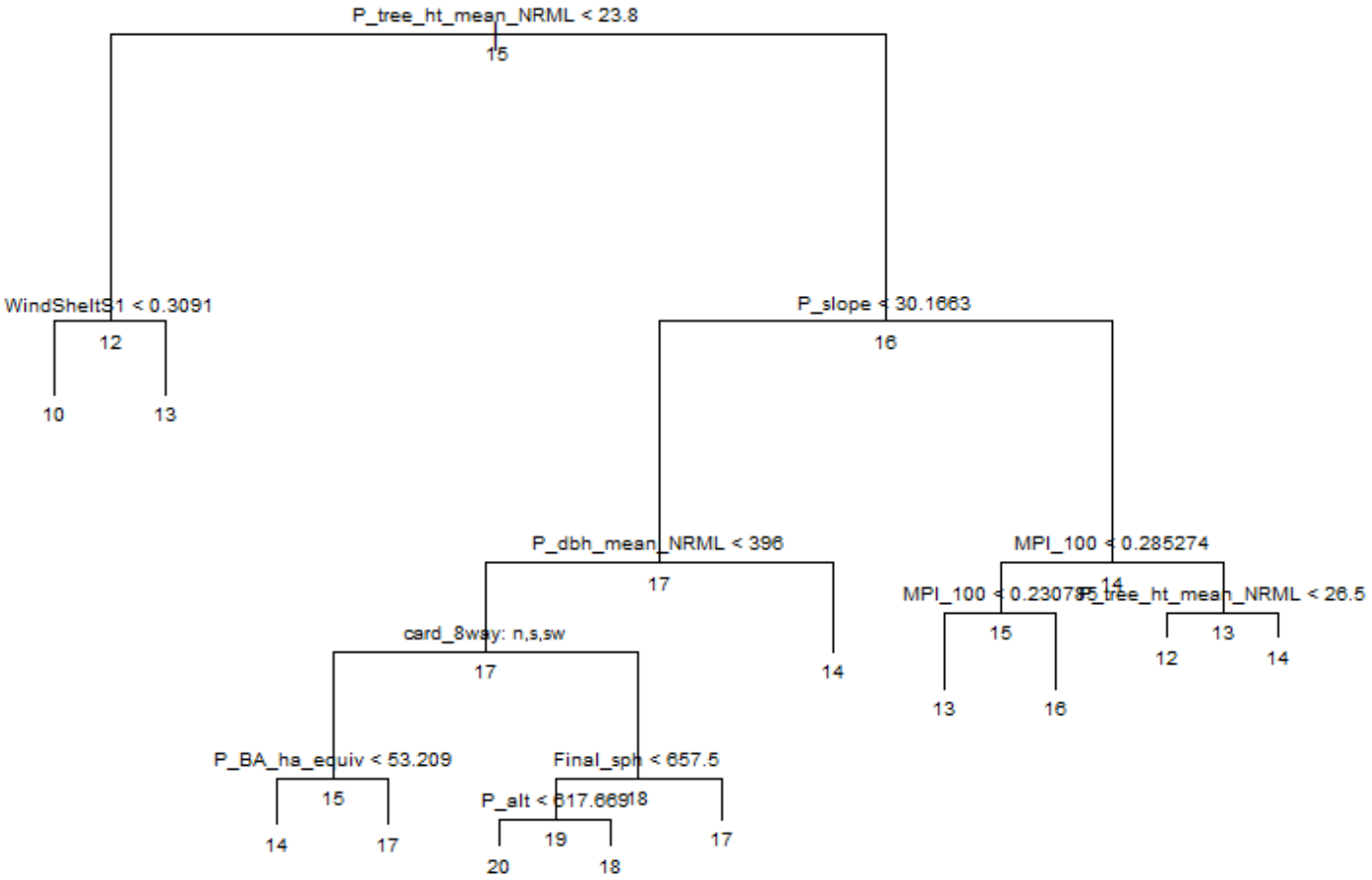

Figure 6-79: CART for response variable P_tree_ht_mean_BRKN and explanatory variables, for Douglas-fir.

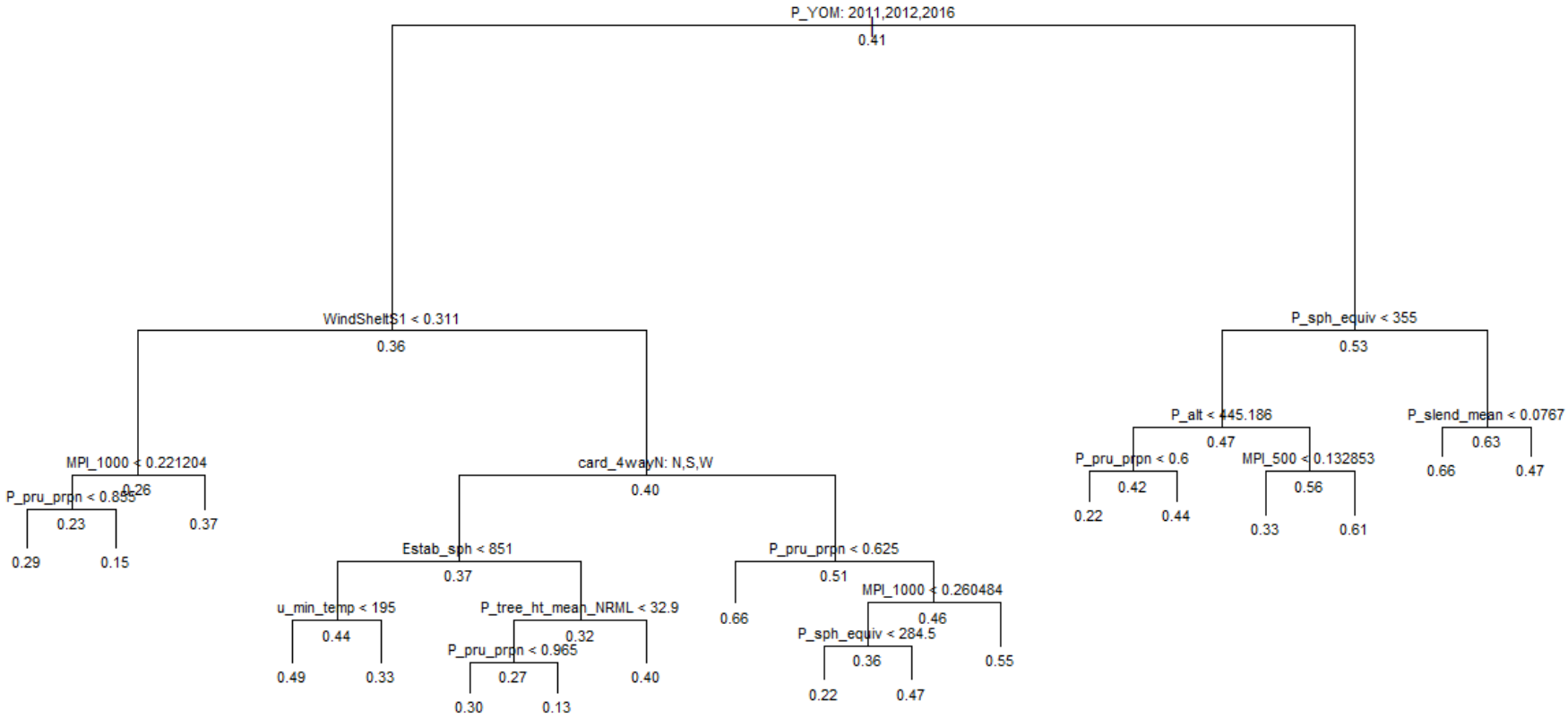

Figure 6-80: CART for response variable Tops_prpn_DAM and explanatory variables, for radiata pine with all plots included.

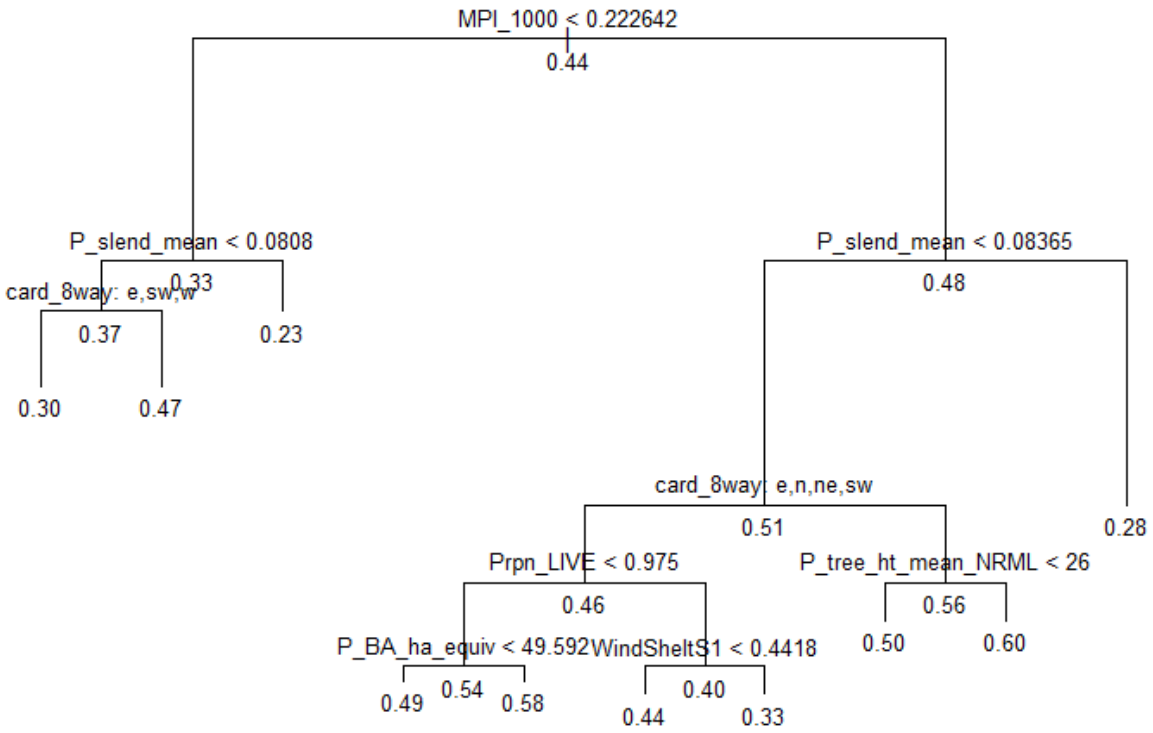

Figure 6-81: CARTs for response variable Tops_prpn_DAM and explanatory variables, for Douglas-fir.

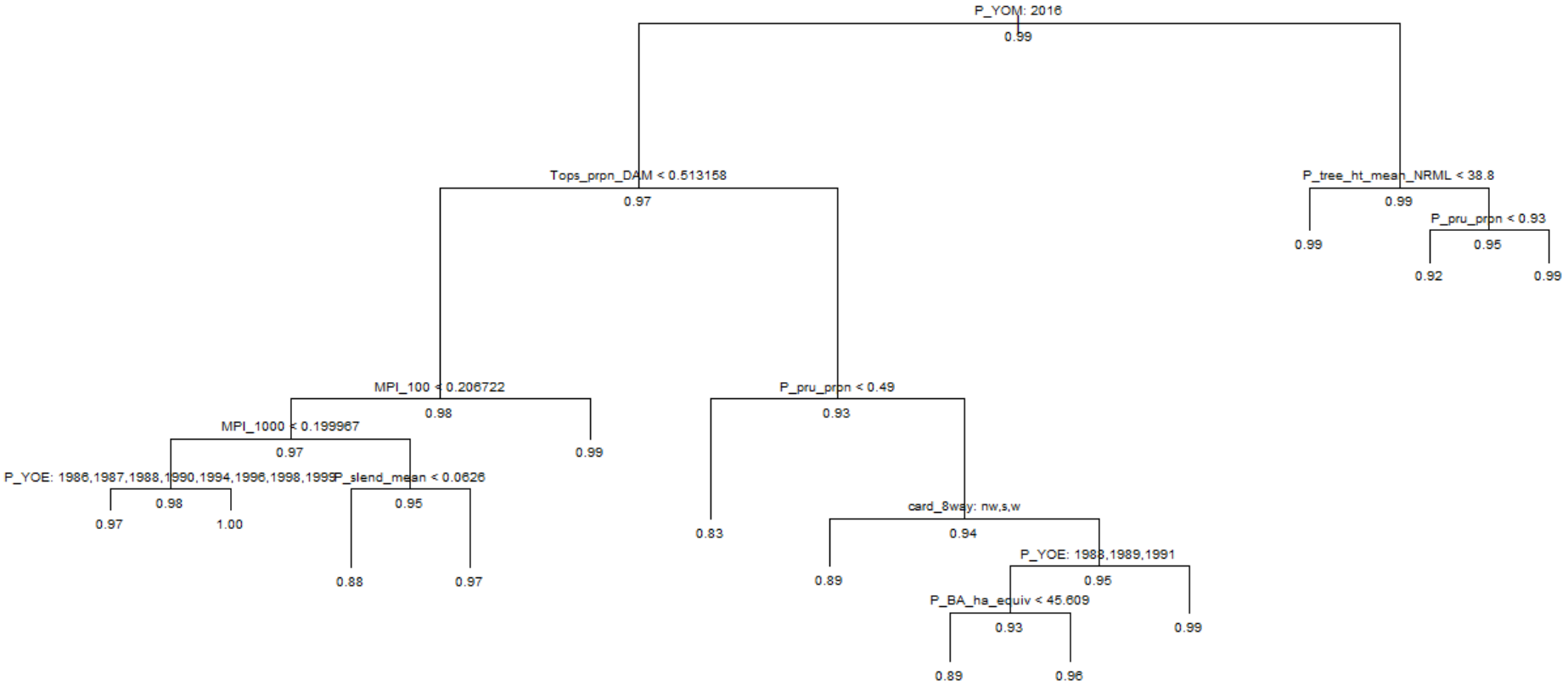

Figure 6-82: CARTs for response variable Prpn_LIVE and explanatory variables, for radiata pine.

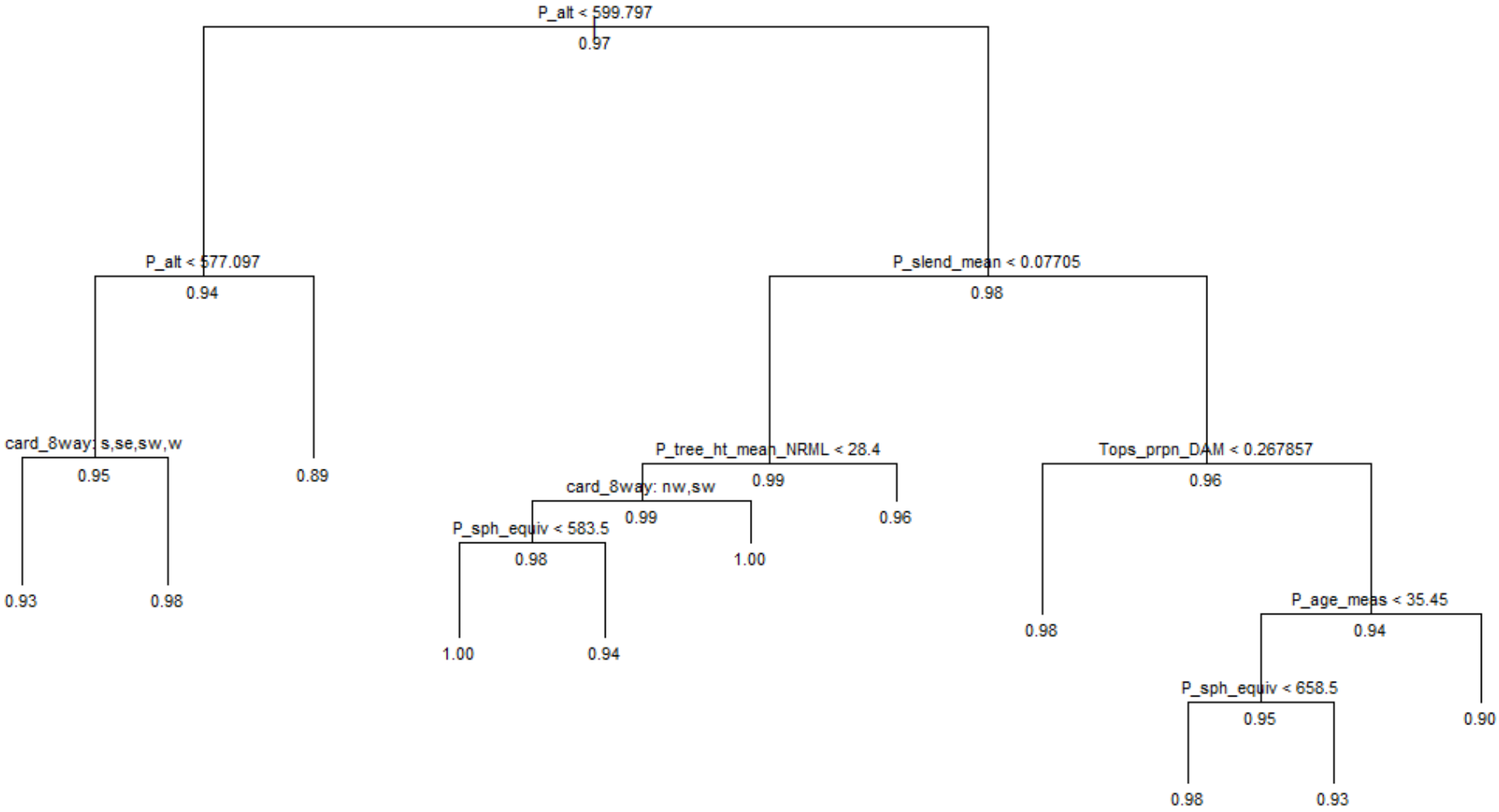

Figure 6-83: CARTs for response variable Prpn_LIVE and explanatory variables, for Douglas-fir.

6.6 Logistic regressions *without* mixed effects

These are the full results for the models that were used to provide the comparison 'R^2 without mixed effects' for logistic regression models. Note: the normally-distributed continuous variable *P_tree_ht_mean_BRKN* does not have an OLRE, so behaves correctly for the marginal and conditional R^2 calculation, and no models of *P_tree_ht_mean_BRKN* appear in this section.

Table 6-10: logistic regression model without mixed effects for radiata pine Tops_prpn_DAM, for all plots.

model type	**logistic regression**					
model fitting	variable	coefficient	std. error	p-value		
	(intercept)	-0.027	0.058	0.6371		
	card_4wayN – N	-0.186	0.070	0.0080		
	card_4wayN – S	-0.340	0.099	0.0006		
	card_4wayN – W	-0.663	0.085	<0.0001		
	MPI_1000_c	0.239	0.029	<0.0001		
	P_thinned	0.613	0.091	<0.0001		
	Prpn_LIVE_c	-0.131	0.024	<0.0001		

fit statistics	R^2 (McFadden's pseudo)	mean	MAPE	bias: intercept	bias: slope	dispersion scale factor
	0.113	0.444	17.6	0.390	0.124	1.72

autocorrelation of residuals	Moran's I observed	Moran's I expected	p-value	autocorrelation present
	0.0397	-0.0020	<0.0001	yes

model testing					
fit statistics	R^2	mean	MAPE	bias: intercept	bias: slope
	0.081	0.453	16.5	0.403	0.116

Table 6-11: logistic regression model without mixed effects for radiata pine Tops_prpn_DAM, only plots with all tops assessed.

model type	**logistic regression**					
model fitting	variable	coefficient	std. error	p-value		
	Intercept	-0.382	0.060	<0.0001		
	P_sph_equiv_cs	0.317	0.043	<0.0001		
	P_Fk_1_prpn_cs	-0.086	0.045	0.0547		
	card_4wayNENW	-0.180	0.084	0.0322		
	card_4wayNESE	0.158	0.148	0.2859		
	card_4wayNESW	-0.339	0.117	0.0037		
	MPI_200_cs	0.298	0.041	<0.0001		
	P_dbh_mean_NRML_cs	0.279	0.052	<0.0001		
	Prpn_LIVE_cs	-0.257	0.040	<0.0001		

fit statistics	R^2 (McFadden's pseudo)	mean	MAPE	bias: intercept	bias: slope	dispersion scale factor
	0.382	0.388	11.2	0.237	0.392	1.96

autocorrelation of residuals	Moran's I observed	Moran's I expected	p-value	autocorrelation present
	0.0133	-0.0061	0.1362	no

model testing					
fit statistics	R^2	mean	MAPE	bias: intercept	bias: slope
	0.387	0.374	14.2	0.261	0.295

Table 6-12: logistic regression model without mixed effects for radiata pine plot Tops_prpn_DAM, all tops, manual hurdle model step 2.

model type	**logistic regression**			
model fitting	variable	coefficient	std. error	p-value
	Intercept	0.112	0.058	0.0517
	P_sph_equiv_cs	0.227	0.034	<0.0001
	P_alt_cs	0.076	0.029	<0.0001
	card_4wayNN	-0.350	0.072	<0.0001
	card_4wayNS	-0.246	0.101	0.0153
	card_4wayNW	-0.684	0.088	<0.0001
	MPI_500_cs	0.254	0.030	<0.0001
	u_wind_tim_cs	-0.110	0.040	<0.0001
	P_dbh_mean_NRML_cs	0.297	0.050	<0.0001
	Prpn_LIVE_cs	-0.100	0.025	<0.0001

fit statistics	R^2 (McFadden's pseudo)	mean	MAPE	bias: intercept	bias: slope	dispersion scale factor
	0.135	0.457	14.0	0.358	0.212	1.6

autocorrelation of residuals	Moran's I observed	Moran's I expected	p-value	autocorrelation present
	0.0384	-0.0022	<0.0001	yes

model testing	not undertaken, as testing would be on combined steps of hurdle model; this is a single step

Table 6-13: logistic regression model without mixed effects for Douglas-fir Tops_prpn_DAM, for all plots.

model type	**logistic regression**			
model fitting	variable	coefficient	std. error	p-value
	Intercept	-0.54673	0.14030	<0.0001
	P_BA_ha_equiv_cs	-0.14515	0.07504	0.0531
	card_4wayNENW	-0.65258	0.22669	0.0040
	card_4wayNESE	-0.26573	0.18256	0.1455
	card_4wayNESW	-0.81656	0.19965	<0.0001
	Prpn_LIVE_cs	-0.28162	0.07456	0.0002

fit statistics	R^2 (McFadden's pseudo)	mean	MAPE	bias: intercept	bias: slope	dispersion scale factor
	0.114	0.272	18.9	0.246	0.126	1.38

autocorrelation of residuals	Moran's I observed	Moran's I expected	p-value	autocorrelation present
	0.0981	-0.0038	<0.0001	yes

model testing					
fit statistics	R^2	mean	MAPE	bias: intercept	bias: slope
	0.119	0.306	18.7	0.252	0.196

Table 6-14: logistic regression model without mixed effects for radiata pine Prpn_LIVE.

model type	**logistic regression**			
model fitting	variable	coefficient	std. error	p-value
	Intercept	4.476	0.149	<0.0001
	P_sph_equiv_cs	-0.244	0.065	0.0002
	P_alt_cs	0.232	0.077	0.0026
	card_4wayNENW	-0.266	0.176	0.1303
	card_4wayNESE	1.257	0.376	0.0008
	card_4wayNESW	-0.370	0.253	0.1441
	P_pru_prpn_cs	0.235	0.081	0.0038
	Estab_sph_cs	0.369	0.091	<0.0001
	u_wind_tim_cs	-0.616	0.105	<0.0001
	Tops_prpn_DAM_cs	-0.296	0.090	<0.0001

fit statistics	R^2 (McFadden's pseudo)	mean	MAPE	bias: intercept	bias: slope	dispersion scale factor
	0.185		2.2	0.816	0.170	1.427

autocorrelation of residuals	Moran's I observed	Moran's I expected	p-value	autocorrelation present
	-0.0030	-0.0020	0.8312	no

model testing					
fit statistics	R^2	mean	MAPE	bias: intercept	bias: slope
	0.117		2.2	0.870	0.118

Table 6-15: logistic regression model without mixed effects for Douglas-fir Prpn_LIVE.

model type	**logistic regression**			
model fitting	variable	coefficient	std. error	p-value
	Intercept	3.648	0.101	<0.0001
	P_sph_equiv_cs	-0.348	0.088	<0.0001
	P_tree_ht_mean_NRML_cs	-0.424	0.088	<0.0001
	Tops_prpn_DAM_cs	-0.296	0.081	<0.0001

fit statistics	R^2 (McFadden's pseudo)	mean	MAPE	bias: intercept	bias: slope	dispersion scale factor
	0.093	0.971	3.1	0.822	0.153	1.025

autocorrelation of residuals	Moran's I observed	Moran's I expected	p-value	autocorrelation present
	-0.0086	-0.0038	0.6369	no

model testing					
fit statistics	R^2	mean	MAPE	bias: intercept	bias: slope
	0.110	0.965	4.6	0.863	0.107

6.7 Random forest outcomes not presented in Results

For every combination of response variable and species modelled, three random forest models were created, with the variable choices 1) all variables included, 2) top ten variables by explanatory power included, as taken from the results for all variables included, and 3) variables that area analogous to those used in the best regression model for the same response variable/species combination. In all cases, type 2) gave the most explanatory power. For completeness, Table 3-5, below, presents the results for types 1) and 3).

Table 6-16: statistics for random forest models performance on fitting data for models not included in Results.

per regression: variables as for the regression for the same response variable.

Response variable	Species	Model variant	variables in model	vars at split	R^2	MAPE	bias: int.	bias: slope
P_tree_ht_ mean_BRKN	radiata pine	all variables	P_age_meas, P_BA_ha_equiv, P_sph_equiv, P_slend_mean, P_Fk_1_prpn, P_pruned, P_pru_prpn, P_pru_ht, card_4wayN, card_4wayNE, card_8way, P_slope, P_alt , MPI_100, MPI_200 , MPI_500, MPI_1000, MPI_2000, P_thinned, Estab_sph, u_wind_tim, u_rain, u_min_temp, u_mint_rain, u_air_pr, u_rain_wind_tim, P_dbh_mean_NRML, P_tree_ht_mean_NRML, Prpn_LIVE, Tops_prpn_DAM, P_YOE, P_YOM, P_stand	6	0.422	25.8	9.427	0.357
		per regression	P_BA_ha_equiv, P_sph_equiv, P_pru_prpn, P_alt, u_wind_tim, P_YOE, P_YOM	3	0.413	25.2	9.225	0.370
	Douglas-fir	all variables	As for radiata pine, but omitting P_pruned, P_pru_prpn, P_pru_ht	6	0.148	25.6	13.119	0.144
		analogue of regression	P_thinned, P_tree_ht_mean_NRML, Tops_prpn_DAM, P_stand	2	0.167	25.0	13.227	0.126
Tops_prpn_ DAM	radiata pine, all plots	all variables	P_age_meas, P_BA_ha_equiv, P_sph_equiv, P_slend_mean, P_Fk_1_prpn, P_pruned, P_pru_prpn, P_pru_ht, card_4wayN, card_4wayNE, card_8way, P_slope, P_alt, MPI_100, MPI_200, MPI_500, MPI_1000, MPI_2000, WindSheltS1, WindSheltNE1, P_thinned, Estab_sph, u_wind_tim, u_rain, u_min_temp, u_mint_rain, u_air_pr, u_rain_wind_tim, P_dbh_mean_NRML, P_tree_ht_mean_NRML, Prpn_LIVE, P_YOE, P_YOM, P_stand	6	0.267	15.9	0.353	0.193
		per regression	card_4wayN, MPI_1000, P_thinned, Prpn_LIVE, P_YOM, P_YOE	3	0.210	16.1	0.337	0.224
	radiata pine, plots with all tops assessed	all variables	P_age_meas, P_BA_ha_equiv, P_sph_equiv, P_slend_mean, P_Fk_1_prpn, P_pruned, P_pru_prpn, P_pru_ht, card_4wayN, card_4wayNE, card_8way, P_slope, P_alt, MPI_100, MPI_200, MPI_500, MPI_1000, MPI_2000, WindSheltS1, WindSheltNE1, P_thinned, Estab_sph, u_wind_tim, u_rain, u_min_temp, u_mint_rain, u_air_pr, u_rain_wind_tim, P_dbh_mean_NRML, P_tree_ht_mean_NRML, Prpn_LIVE	6	0.223	13.1	0.302	0.221
		per regression	P_sph_equiv, card_4wayN, MPI_200, P_dbh_mean_NRML, Prpn_LIVE, P_YOE	2	0.165	13.2	0.304	0.226
	Douglas-fir	all variables	As for radiata pine, but omitting P_pruned, P_pru_prpn, P_pru_ht	6	0.200	16.9	0.178	0.175
		per regression	P_BA_ha_equiv, P_sph_equiv, card_4wayNE, Prpn_LIVE, P_YOE	2	0.112	17.4	0.185	0.142
Prpn_LIVE	radiata pine	all variables	P_age_meas, P_BA_ha_equiv, P_sph_equiv, P_slend_mean, P_Fk_1_prpn, P_pruned, P_pru_prpn, P_pru_ht, card_4wayN, card_4wayNE, card_8way, P_slope, P_alt, MPI_100, MPI_200, MPI_500, MPI_1000, MPI_2000 , WindSheltS1, WindSheltNE1, P_thinned, Estab_sph, u_wind_tim, u_rain, u_min_temp, u_mint_rain, u_air_pr, u_rain_wind_tim, P_dbh_mean_NRML, P_tree_ht_mean_NRML, Tops_prpn_DAM, P_YOE, P_YOM, P_stand	6	0.148	2.2	0.860	0.126
		per regression	P_sph_equiv, P_slend_mean, card_4wayNE, P_pru_prpn, Estab_sph, u_wind_tim, P_dbh_mean_NRML, Tops_prpn_DAM, P_YOE	3	0.161	2.2	0.840	0.146
	Douglas-fir	all variables	As for radiata pine, but omitting P_pruned, P_pru_prpn, P_pru_ht	6	0.073	3.4	0.902	0.069
		per regression	P_sph_equiv, Estab_sph, P_tree_ht_mean_NRML, Tops_prpn_DAM	2	0.038	3.4	0.904	0.067

6.8 Photographs of wind damage at Geraldine Forest

These photographs show quite severe wind damage, including a mixture of breakage and windthrow. Areas affected to this degree would be mapped out of stands and removed from the tally of the productive area of the forest. However, the manner in which trees snap (near the bottom of the green crown, or below it), is the same as for more scattered cases of broken standing trees, as studied in this research. Photographs by Phil Taylor, Port Blakely Ltd.

7 References

Achim, A., Ruel, J. C., Gardiner, B. A., Laflamme, G., & Meunier, S. (2005). Modelling the vulnerability of balsam fir forests to wind damage. *Forest Ecology and Management, 204*(1), 37-52. https://doi.org/10.1016/j.foreco.2004.07.072

Albrecht, A., Hanewinkel, M., Bauhus, J., & Kohnle, U. (2012). How does silviculture affect storm damage in forests of south-western Germany? Results from empirical modeling based on long-term observations. *European Journal of Forest Research, 131*(1), 229-247. https://doi.org/10.1007/s10342-010-0432-x

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: A review. *International Journal of Advances in Soft Computing and its Applications, 7*(3), 176-204.

Aszalós, R., Somodi, I., Kenderes, K., Ruff, J., Czúcz, B., & Standovár, T. (2012). Accurate prediction of ice disturbance in European deciduous forests with generalized linear models: a comparison of field-based and airborne-based approaches. *European Journal of Forest Research, 131*(6), 1905-1915. https://doi.org/10.1007/s10342-012-0641-6

Aubrey, D. P., Coleman, M. D., & Coyle, D. R. (2007). Ice damage in loblolly pine: Understanding the factors that influence susceptibility. *Forest Science, 53*(5), 580-589.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1 - 48. https://doi.org/https://www.jstatsoft.org/article/view/v067i01

Bennett, M. P. (2002). *The analysis of the preliminary results obtained from the Otago Coast Forest windthrow experiment: a dissertation submitted in partial fulfillment of the requirements for the degree of Bachelor of Forestry Science, School of Forestry, University of Canterbury, Christchurch, New Zealand* (Undergraduate dissertation, University of Canterbury, Christchurch, New Zealand).

Blennow, K., & Olofsson, E. (2008). The probability of wind damage in forestry under a changed wind climate. *Climatic Change, 87*(3), 347-360. https://doi.org/10.1007/s10584-007-9290-z

Bolker, B. (2019). GLMM FAQ. Retrieved from https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#overdispersion

Chapman, L. (2000). Assessing topographic exposure. *Meteorological applications, 7*(4), 335-340. https://doi.org/10.1017/S1350482700001729

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., . . . Boehner, J. (2015). System for Automated Geoscientific Analyses (SAGA) (Version 2.1.4). Retrieved from http://www.geosci-model-dev.net/8/1991/2015/gmd-8-1991-2015.html https://doi.org/doi:10.5194/gmd-8-1991-2015

Cucchi, V., Meredieu, C., Stokes, A., de Coligny, F., Suarez, J., & Gardiner, B. A. (2005). Modelling the windthrow risk for simulated forest stands of Maritime pine (Pinus pinaster Ait.). *Forest Ecology and Management, 213*(1), 184-196. https://doi.org/10.1016/j.foreco.2005.03.019

Díaz-Yáñez, O., Mola-Yudego, B., & González-Olabarria, J. R. (2019). Modelling damage occurrence by snow and wind in forest ecosystems. *Ecological Modelling, 408*(108741). https://doi.org/10.1016/j.ecolmodel.2019.108741

Díaz-Yáñez, O., Mola-Yudego, B., González-Olabarria, J. R., & Pukkala, T. (2017). How does forest composition and structure affect the stability against wind and snow? *Forest Ecology and Management, 401*, 215-222. https://doi.org/10.1016/j.foreco.2017.06.054

Dobbertin, M. (2002). Influence of stand structure and site factors on wind damage comparing the storms Vivian and Lothar. *Forest Snow and Landscape Research, 77*(1-2), 187-205.

Elie, J.-G., & Ruel, J.-C. (2005). Windthrow hazard modelling in boreal forests of black spruce and jack pine. *Canadian Journal of Forest Research, 35*(11), 2655-2663. https://doi.org/10.1139/x05-189

ESRI. (2017). ArcGIS Desktop: Release 10.6. . Redlands, California, U.S.A.: Environmental Systems Research Institute.
Everham, E. M. I., & Nicholas, V. L. B. (1996). Forest Damage and Recovery from Catastrophic Wind. *Botanical Review, 62*(2), 113-185. https://doi.org/10.1007/BF02857920
Fletcher, D., MacKenzie, D., & Villouta, E. (2005). Modelling skewed data with many zeros: A simple approach combining ordinary and logistic regression. *Environmental and Ecological Statistics, 12*(1), 45-54. https://doi.org/10.1007/s10651-005-6817-1
Fridman, J., Valinger, E., & Sveriges, l. (1998). Modelling probability of snow and wind damage using tree, stand, and site characteristics from Pinus sylvestris sample plots. *Scandinavian Journal of Forest Research, 13*(1-4), 348-356. https://doi.org/10.1080/02827589809382994
Goulding, C. J. (2005). Measurement of Trees. In M. Colley (Ed.), *Forestry Handbook* (pp. 145 - 148). Christchurch, New Zealand: New Zealand Institute of Forestry (Inc).
Hanewinkel, M., Breidenbach, J., Neeff, T., & Kublin, E. (2008). Seventy-seven years of natural disturbances in a mountain forest area the influence of storm, snow, and insect damage analysed with a long-term time series. *Canadian Journal of Forest Research, 38*(8), 2249-2261. https://doi.org/10.1139/X08-070
Hanewinkel, M., Kuhn, T., Bugmann, H., Lanz, A., & Brang, P. (2014). Vulnerability of uneven-aged forests to storm damage. *Forestry: An International Journal of Forest Research, 87*(4), 525-534. https://doi.org/10.1093/forestry/cpu008
Hannah, P., Palutikof, J. P., & Quine, C. P. (1995). Predicting windspeeds for forest areas in complex terrain. In M. P. Coutts & J. Grace (Eds.), *Wind and trees* (pp. 113–129). Cambridge, U.K.: Cambridge University Press.
Hansen, W. J., & Cranson, J. (2016). Spatial Analysis of Forest Damage in Central Massachusetts Resulting from the December 2008 Ice Storm. *Northeastern Naturalist, 23*(3), 378-394. https://doi.org/10.1656/045.023.0306
Harrison, X. A. (2015, 07/21). A comparison of observation-level random effect and Beta-Binomial models for modeling overdispersion in Binomial data in ecology & evolution. *PeerJ, 3*, e1114. https://doi.org/10.7717/peerj.1114
Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., . . . Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ, 2018*(5), e4794-e4794. https://doi.org/10.7717/peerj.4794
Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics, 7*(3), 355-373. https://doi.org/10.1093/biostatistics/kxj011
Jalkanen, A., & Mattila, U. (2000). Logistic regression models for wind and snow damage in northern Finland based on the National Forest Inventory data. *Forest Ecology and Management, 135*(1), 315-330. https://doi.org/10.1016/S0378-1127(00)00289-9
Knowles, R. L., & Paton, V. J. (1989). The Effect of Final-Crop Stocking on Wind Damage at Tikitere. In A. Somerville, S. Wakelin, & L. Whitehouse (Eds.), *Workshops on Wind damage in New Zealand Exotic Forests* (pp. 37 - 37). Rotorua, New Zealand New Zealand Forest Research Institute.
Krejci, L., Kolejka, J., Vozenilek, V., & Machar, I. (2018). Application of GIS to Empirical Windthrow Risk Model in Mountain Forested Landscapes. *Forests, 9*(2), 96. https://doi.org/10.3390/f9020096
Land Information New Zealand. (2008). *LINZS25002: Standard for New Zealand Geodetic Datum 2000 Projections: version 2*. Wellington, New Zealand: Office of the Surveyor-General, Land Information New Zealand.
Landcare Research New Zealand Ltd. (2015). *New Zealand Land Cover Database*. New Zealand: Landcare Research New Zealand Ltd. Retrieved from https://lris.scinfo.org.nz/
Lanquaye-Opoku, N., & Mitchell, S. J. (2005). Portability of stand-level empirical windthrow risk models. *Forest Ecology and Management, 216*(1), 134-148. https://doi.org/10.1016/j.foreco.2005.05.032

Ledgard, D. R. (1982). *A qualitative and quantitative examination of alternative silvicultural regimes in the South Canterbury foothills with reference to the evaluation of climatic factors that may affect these regimes: a dissertation submitted in partial fulfilment of the requirements for the degree of Bachelor of Forestry Science in the University of Canterbury* (Undergraduate dissertation, University of Cantebury, Christchurch, New Zealand).

Lefcheck, J. S. (2016). Piecewise structural equation modeling in R for ecology, evolution, and systematics. *Methods in Ecology and Evolution, 7*(5), 573-579. https://doi.org/10.1111/2041-210X.12512

Lindemann, J. D., & Baker, W. L. (2002). Using GIS to analyse a severe forest blowdown in the Southern Rocky Mountains. *International Journal of Geographical Information Science, 16*(4), 377-399. https://doi.org/10.1080/13658810210136069

Maalouf, M., & Siddiqi, M. (2014). Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems, 59*, 142-148. https://doi.org/10.1016/j.knosys.2014.01.012

Martín-Alcón, S., González-Olabarria, J. R., & Coll, L. (2010). Wind and snow damage in the Pyrenees pine forests: Effect of stand attributes and location. *Silva Fennica, 44*(3), 399-410. https://doi.org/10.14214/sf.138

Martin, T. J., & Ogden, J. (2006). Wind damage and response in New Zealand forests: a review. *New Zealand Journal of Ecology, 30*(3), 295-310.

Ministry for Primary Industries. (2018). *National Exotic Forest Description as at 1 April 2018*. Wellington, New Zealand: Ministry for Primary Industries

Mitchell, S. J. (2013). Wind as a natural disturbance agent in forests: A synthesis. *Forestry, 86*(2), 147-157. https://doi.org/10.1093/forestry/cps058

Mitchell, S. J., Hailemariam, T., & Kulis, Y. (2001). Empirical modeling of cutblock edge windthrow risk on Vancouver Island, Canada, using stand level information. *Forest Ecology and Management, 154*(1), 117-130. https://doi.org/10.1016/S0378-1127(00)00620-4

Moore, J. R., & Gardiner, B. (2001). Relative windfirmness of New Zealand-grown Pinus radiata and Douglas-fir: A preliminary investigation. *New Zealand Journal of Forestry Science, 31*(2), 208-223.

Moore, J. R., Manley, B. R., Park, D., & Scarrott, C. J. (2013). Quantification of wind damage to New Zealand's planted forests. *Forestry: An International Journal of Forest Research, 86*(2), 173-183. https://doi.org/10.1093/forestry/cps076

Moore, J. R., & Quine, C. P. (2000). A comparison of the relative risk of wind damage to planted forests in Border Forest Park, Great Britain, and the Central North Island, New Zealand. *Forest Ecology and Management, 135*(1), 345-353. https://doi.org/10.1016/S0378-1127(00)00292-9

Moore, J. R., & Somerville, A. (1998). Assessing the risk of wind damage to plantation forests in New Zealand. *New Zealand forestry, 43*(1), 25-29.

Munishi, P. K. T., & Chamshama, S. A. O. (1994). A study of wind damage on Pinus patula stands in Southern Tanzania. *Forest Ecology and Management, 63*(1), 13-21. https://doi.org/10.1016/0378-1127(94)90244-5

Nakagawa, S., Schielzeth, H., & O'Hara, R. B. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133-142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

National Institute of Water and Atmospheric Research. (2019a). Overview of New Zealand's climate. Retrieved 14/12/2019, 2019, from https://niwa.co.nz/education-and-training/schools/resources/climate/overview

National Institute of Water and Atmospheric Research. (2019b). Virtual Climate Station data and products. Retrieved 17 July, 2019, from https://www.niwa.co.nz/climate/our-services/virtual-climate-stations

Nixon, C., Gamperle, D., Pambudi, D., & Clough, P. (2017, March 2017). *Plantation forestry statistics: Contribution of forestry to New Zealand*. New Zealand: New Zealand Institute of Economic Research.

Olsen, P. F. (1989). Wind Risk - a Consultant's Perspective. In A. Somerville, S. Wakelin, & L. Whitehouse (Eds.), *Workshops on Wind damage in New Zealand Exotic Forests* (p. 26). Rotorua, New Zealand New Zealand Forest Research Institute.

Päätalo, M.-L. (2000). Risk of Snow Damage in Unmanaged and Managed Stands of Scots Pine, Norway Spruce and Birch. *Scandinavian Journal of Forest Research, 15*(5), 530-541. https://doi.org/10.1080/028275800750173474

Paradis , E., & Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics, 35*, 526-528.

Park, D. (2009). *Documentation and analysis of catastrophic wind damage in New Zealand radiata pine plantations: a dissertation submitted in partial fulfilment of the requirements for the degree of Bachelor of Forestry Science with Honours, New Zealand School of Forestry, University of Canterbury, Christchurch, New Zealand* (Undergraduate dissertation, University of Canterbury, Christchurch, New Zealand).

Peltola, H., Väisänen, H., Kellomäki, S., & Ikonen, V. P. (1999). A mechanistic model for assessing the risk of wind and snow damage to single trees and stands of Scots pine, Norway spruce, and birch. *Canadian Journal of Forest Research, 29*(6), 647-661. https://doi.org/10.1139/x99-029

Quine, C. P. (1995). Assessing the risk of wind damage to forests: practice and pitfalls. In J. Grace & M. P. Coutts (Eds.), *Wind and Trees* (pp. 379-403). Cambridge: Cambridge University Press.

R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rebertus, A. J., Kitzberger, T., Veblen, T. T., & Roovers, L. M. (1997). Blowdown History and Landscape Patterns in the Andes of Tierra del Fuego, Argentina. *Ecology, 78*(3), 678-692. https://doi.org/10.1890/0012-9658(1997)078[0678:BHALPI]2.0.CO;2

Rifai, S. W., Urquiza Muñoz, J. D., Negrón-Juárez, R. I., Ramírez Arévalo, F. R., Tello-Espinoza, R., Vanderwel, M. C., . . . Bohlman, S. A. (2016). Landscape-scale consequences of differential tree mortality from catastrophic wind disturbance in the Amazon. *Ecological Applications, 26*(7), 2225-2237. https://doi.org/10.1002/eap.1368

Ruel, J.-C. (1995). Understanding windthrow - Silvicultural implications. *Forestry Chronicle, 71*(4), 434-445. https://doi.org/10.5558/tfc71434-4

Schindler, D., Bauhus, J., & Mayer, H. (2012). Wind effects on trees. *European Journal of Forest Research, 131*(1), 159-163. https://doi.org/10.1007/s10342-011-0582-5

Schmidt, M., Hanewinkel, M., Kändler, G., Kublin, E., & Kohnle, U. (2010). An inventory-based approach for modeling singletree storm damage - experiences with the winter storm of 1999 in southwestern Germany. *Canadian Journal of Forest Research, 40*(8), 1636-1652. https://doi.org/10.1139/X10-099

Scott, R. E., & Mitchell, S. J. (2005). Empirical modelling of windthrow risk in partially harvested stands using tree, neighbourhood, and stand attributes. *Forest Ecology and Management, 218*(1), 193-209. https://doi.org/10.1016/j.foreco.2005.07.012

Somerville, A. (1995). Wind damage to New Zealand State plantation forests. In J. Grace & M. P. Coutts (Eds.), *Wind and Trees* (pp. 460-467). Cambridge: Cambridge University Press.

Somerville, A., Wakelin, S., & Whitehouse, L. (Eds.). (1989). *Workshop on wind damage in New Zealand exotic forests* (Vol. 146). Rotorua, New Zealand: New Zealand Forest Research Institute.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*(1), 307-307. https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*(1), 25-25. https://doi.org/10.1186/1471-2105-8-25

Stueve, K. M., Lafon, C. W., & Isaacs, R. E. (2007). Spatial patterns of ice storm disturbance on a forested landscape in the Appalachian Mountains, Virginia. *Area (London 1969), 39*(1), 20-30. https://doi.org/10.1111/j.1475-4762.2007.00722.x

Turner, R. M. (1989). Canterbury Plains and Lake Taupo Forests. In A. Somerville, S. Wakelin, & L. Whitehouse (Eds.), *Workshops on Wind damage in New Zealand Exotic Forests* (pp. 10 - 12). Rotorua, New Zealand New Zealand Forest Research Institute.

University of Otago School of Surveying. (2011). NZSoSDEM v1.0

Usbeck, T., Waldner, P., Dobbertin, M., Ginzler, C., Hoffmann, C., Sutter, F., . . . Rebetez, M. (2012). Relating remotely sensed forest damage data to wind data: storms Lothar (1999) and Vivian (1990) in Switzerland. *Theoretical and Applied Climatology, 108*(3), 451-462. https://doi.org/10.1007/s00704-011-0526-5

Valinger, E., & Fridman, J. (1999). Models to Assess the Risk of Snow and Wind Damage in Pine, Spruce, and Birch Forests in Sweden. *Environmental Management, 24*(2), 209-217. https://doi.org/10.1007/s002679900227

Valinger, E., & Fridman, J. (2011). Factors affecting the probability of windthrow at stand level as a result of Gudrun winter storm in southern Sweden. *Forest Ecology and Management, 262*(3), 398-403. https://doi.org/10.1016/j.foreco.2011.04.004

Valinger, E., & Pettersson, N. (1996). Wind and snow damage in a thinning and fertilization experiment in Picea abies in southern Sweden. *Forestry, 69*(1), 25-33. https://doi.org/10.1093/forestry/69.1.25

Veblen, T. T., Kulakowski, D., Eisenhart, K. S., & Baker, W. L. (2001). Subalpine forest damage from a severe windstorm in northern Colorado. *Canadian Journal of Forest Research, 31*(12), 2089-2097. https://doi.org/10.1139/cjfr-31-12-2089

Wallentin, C., & Nilsson, U. (2014). Storm and snow damage in a Norway spruce thinning experiment in southern Sweden. *FORESTRY, 87*(2), 229-238. https://doi.org/10.1093/forestry/cpt046

Wang, F., & Xu, Y. J. (2009). Hurricane Katrina-induced forest damage in relation to ecological factors at landscape scale. *Environmental Monitoring and Assessment, 156*(1-4), 491-507. https://doi.org/10.1007/s10661-008-0500-6

Wei, T., & Simco, V. (2017). R package "corrplot": Visualisation of a Correlation Matrix (Version 0.84). Retrieved from https://github.com/taiyun/corrplot

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. New York, U.S.A.: Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Woollons, R. C. (1998). Even-aged stand mortality estimation through a two-step regression process. *Forest Ecology and Management, 105*(1), 189-195. https://doi.org/10.1016/S0378-1127(97)00279-X

Wrathall, S. H. (1989). *Cyclone Bola & Waitahanui forest: a study in catastrophic wind damage.* (Dissertation, University of Canterbury, Christchurch, New Zealand).

Wright, J. A., & Quine, C. P. (1993). The use of a geographical information system to investigate storm damage to trees at Wykeham Forest, North Yorkshire. *Scottish Forestry, 47*(4), 166-174.

Yee, T. W. (2019). VGAM: Vector Generalized Linear and Additive Models. R package version 1.1-1. Retrieved from https://CRAN.R-project.org/package=VGAM

Yokoyama, R., Shirasawa, M., & Pike, R. J. (2002). Visualizing topography by openness: A new application of image processing to digital elevation models. *Photogrammetric Engineering and Remote Sensing, 68*(3), 257-265.

Zeileis, A., & Hothorn, T. (2002). Diagnostic Checking in Regression Relationships: vignette for package lmtest. Retrieved 01/10/2019, 2019, from https://ggplot2.tidyverse.org

Zeng, H., Garcia-Gonzalo, J., Peltola, H., & Kellomäki, S. (2010). The effects of forest structure on the risk of wind damage at a landscape level in a boreal forest ecosystem. *Annals of Forest Science, 67*(1), 111-111. https://doi.org/10.1051/forest/2009090